(2)

ARO-UR Center for

OPTO-ELECTRONIC SYSTEMS RESEARCH

TECHNICAL REPORT

CLASSIFICATION TECHNIQUES FOR QUANTUM-LIMITED AND CLASSICAL-INTENSITY IMAGES

Miles N. Wernick

December 1989

The Institute of Optics
University of Rochester

Prepared for:
U.S. Army Research Office
ATTN: DRXRO-IP-Library
P. O. Box 12211
Research Triangle Park, NC 27709

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARO 24626.129-PH-UIR
6a. NAME OF PERFORMING ORGANIZATION University of Rochester 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION U. S. Army Research Office
6c. ADDRESS (City, State, and ZIP Code) The Institute of Optics Rochester, New York 14627	7b. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U. S. Army Research Office 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
Classification Techniques for Quantum-Limited and Classical-Intensity Images

12. PERSONAL AUTHOR(S)
Miles N. Wernick

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM _____ TO _____	December 1989	139

16. SUPPLEMENTARY NOTATION
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
image classification; pattern recognition, photon-counting, convex hull

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Please see Abstract on Page iv

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL G. Michael Morris	22b. TELEPHONE (Include Area Code) 716-275-5140 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

Classification Techniques for Quantum-Limited and Classical-Intensity Images

by

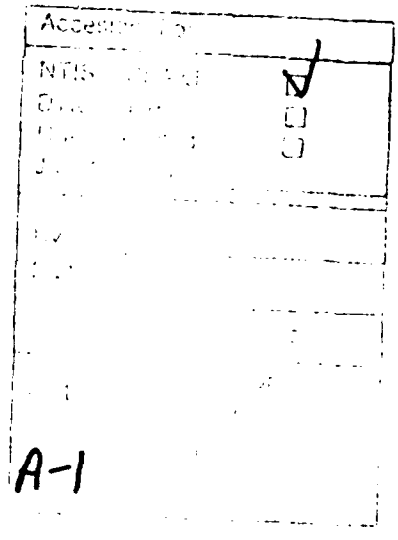
Miles N. Wernick

Submitted in Partial Fulfillment

of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY



Supervised by Dr. G. Michael Morris

The Institute of Optics

University of Rochester

Rochester, NY

1989

Curriculum Vitae

Miles Wernick [redacted]. In 1979 he entered Northwestern University as a National Merit Scholar. In 1983, he graduated from Northwestern with a Bachelor of Arts degree in physics and went on to attend the University of Rochester's Institute of Optics where he pursued studies leading to the present thesis.

Acknowledgments

First, I wish to thank my advisor, Mike Morris, with whom it has been a pleasure to work these past years. He has lent invaluable support in many ways and his racquetball tips are unbeatable. I would also like to acknowledge my friends within our research group. They have provided countless useful technical discussions and incessant inane banter, both of which have made the 4th floor experience a good one. I wish, in particular, to acknowledge Lennart Sääf and Tony Martino for many productive interactions and Tom Isberg for his part in the experimental work described in Section 3.7.1.

I would also like to thank my friends from outside the group, so many of whom are in some unfortunate way connected with the corner table at the Elmwood. In particular, I wish to thank John Conley and Simon Wilkie for many stimulating technical discussions.

Special thanks go to my good friend Hesna Genay, a first-class fıstık, for her advice about Kuhn-Tucker conditions and for her support and patience during this effort. Special thanks also to my family, especially Sherm and Katie, for more things than I could hope to name here. Their help and support began many years before this research project was undertaken.

Finally, I wish to acknowledge the generous financial support of the Army Research Office and the New York State Center for Advanced Optical Technology.

Abstract

Automatic classification of both quantum-limited and classical-intensity images is considered. A new discriminant vector for classical-intensity images is proposed which can also be applied to general pattern recognition problems. The proposed discriminant vector, based on convex analysis, is shown in all cases to correctly distinguish two image classes, if that is, indeed, possible with a linear discriminant. Further, it is shown to be the discriminant vector that maximizes the minimum separation of the values obtained by forming the inner products between it and the input images. Experiments are reported in which the discriminant vector is used to successfully distinguish first two classes, then eight classes of images.

The classification problem is also considered for the case of quantum-limited input images. Quantum-limited images arise as a matter of course in applications such as night vision, low-dose electron microscopy, and radiological imaging. It can also be shown, however, that for reasons of computational efficiency it may be advantageous to use a quantum-limited imaging system as the input to the classifier. The inner product between a quantum-limited image and a discriminant vector is shown to be a Monte Carlo estimator of the corresponding high-light-level inner product. In principle, therefore, any linear classifier can be implemented using the quantum-limited system. In addition, however, new solutions, specifically designed for quantum-limited images are derived using statistical decision theory. These solutions are shown experimentally to provide excellent results. The method is then extended to permit classification of quantum-limited images despite in-plane rotations, and the result is demonstrated experimentally. Finally, the application of the low-light-level solutions to high-light-level situations is considered.

Table of contents

List of figures

List of tables

Chapter 1:
Introduction

Central to activities ranging from medical diagnosis to industrial quality control is the process of interpreting and making decisions based upon information gained through various measurement procedures. Increasingly, imaging technologies of all kinds are relied upon to provide this kind of information. In recent years, an effort has been made to automate the process of extracting information from images and to construct systems that use that information to make decisions.

In industry, for example, optical images, spectra, and surface profile maps are used by inspection, assembly, safety, and control systems. In medical and military applications, sophisticated imaging techniques provide essential data required for decision-making. In the commercial arena, identification of printed and hand-written text permits automatic mail sorting and document processing. Throughout the scientific community, various kinds of spectral information have long been used for measurement and analysis but have only rather recently been interpreted automatically.

Images, as suggested by the above examples, are one- or two-dimensional spatial functions. An important element of many image analysis applications is the ability to automatically classify or categorize these functions. The significance of classifying an image (as opposed to specifically identifying it) derives from the manner in which the image classes are defined. If, for example, images of cars are used to form a class, then a classifier can determine that an object is a car regardless of the

details that distinguish it from other cars. Alternatively, if various images of a single car are used to define a class, then that particular car can be identified despite potentially misleading variations in illumination, perspective, and other such factors.

It is also possible to use the classifier solely for computational advantage in an image recognition task. By grouping images into classes and forming a hierarchical decision tree, the number of decisions required to recognize an image can be greatly reduced. The automatic recognition of printed characters can be achieved, for example, by a series of binary decisions that eliminate subsets of the character set. In this way, N characters (perhaps subject to font variations) can be classified in roughly $\log_2 N$ binary decisions.

Image classification can be achieved in many ways, but most techniques can be grouped into three categories: artificial intelligence (AI) approaches, structural approaches, and correlation-based techniques.[1,2] The literature describing the former two approaches is extensive and is beyond the scope of this thesis. For an introduction to these approaches, the reader is directed to Refs. 3 and 4. This thesis will focus instead on the latter, correlation-based, approach to image classification.

The common feature of correlation-based approaches is that the extraction of information from an image is achieved by forming the cross-correlation or the inner product between the image and some reference function or discriminant vector. This approach is well suited to optical, digital, and hybrid implementations and while it lacks some of the flexibility of the other methods, it generally holds the advantage in speed. In addition, the generality of correlation-based approaches suggests their immediate application to a wide range of problems from object recognition to mass spectral

analysis. Finally, although the human visual system is highly nonlinear, it has been shown that for certain medical diagnosis tasks, human observer performance is correlated with that of a linear classifier.[5] The study of linear classifiers may, therefore, provide useful insights for optimization of medical imaging systems.

An important feature of most correlation-based image classification methods is the vector representation for images. This method of image description is the subject of the following section.

1.1 The vector representation of images

Suppose that an image intensity distribution, $f(x,y)$, is sampled on a rectangular grid of $N_x \times N_y$ ($= N$) picture elements (pixels). The resulting values can be ordered into an N-dimensional vector $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, in which the ith component of $\mathbf{x}$ is the ith sample of the function f. Any image can, in this way, be specified as a point in an N-dimensional vector space. Images described in this way will henceforth be referred to as image vectors.

It is often instructional to consider graphical representations of these vectors. Since drawings of higher-dimensional spaces are impossible, figures based on hypothetical examples of images containing only two pixels will be used to illustrate the principles. Please keep in mind that actual image vectors are composed of thousands of components.

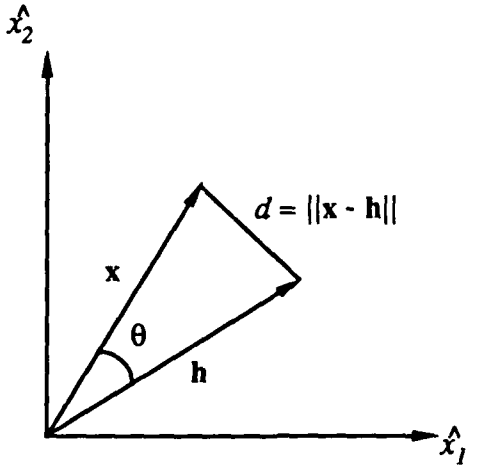


Figure 1.1 Angular and distance measures between an image vector and a template vector.

In the vector representation, one approach to recognizing an image is to compare its vector to a template or reference vector, **h**, by computing a measure of similarity. If the resulting similarity measure exceeds some threshold, then we can say that the input and reference images are the same; hence the input image is "recognized." The most common choices for the similarity measure are *i*) c, the cosine of the angle, θ, between the input and reference vectors, and *ii*) d^2, the squared Euclidean distance between their endpoints (see Fig. 1.1).

The quantity c is given by

$$c = \frac{\mathbf{x} \cdot \mathbf{h}}{\|\mathbf{x}\|\|\mathbf{h}\|} \tag{1.1}$$

or

$$c = \frac{\sum_{i=1}^{N} x_i h_i}{\left[\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} h_i^2\right]^{1/2}} \quad . \tag{1.2}$$

This is simply a discrete form for the output of the normalized matched filter[6] evaluated at the origin.

The quantity d^2 is closely related to the normalized inner product of Eq. (1.1). It is defined in the usual way as

$$d^2 = \sum_{i=1}^{N} (x_i - h_i)^2 \quad , \tag{1.3}$$

which, when expanded, becomes

$$d^2 = -2\mathbf{x} \cdot \mathbf{h} + \|\mathbf{x}\|^2 + \|\mathbf{h}\|^2 \quad . \tag{1.4}$$

At this point, it is useful to note that any scalar multiple of an image vector **x** describes a brighter or darker version the same image. For most purposes, therefore, we wish to make no distinction between image vectors **x** and $k\mathbf{x}$ ($k \in R^+$). When that is indeed the

case, the distance criterion is only meaningful if the input and reference vectors are first normalized. If the vectors are pre-normalized, then c becomes simply

$$c = \mathbf{x} \cdot \mathbf{h} \quad , \tag{1.5}$$

and d^2 becomes

$$d^2 = -2c + 2 \quad . \tag{1.6}$$

Concerning the simple image recognition task, the conclusions of this discussion are: *i*) that the distance and angular similarity measures reduce to inner product calculations, and *ii*) since they are linearly related, they convey equivalent information.

Now let us turn our attention to the subject of this thesis, the problem of sorting or classifying images. Consider, for example, the hypothetical two-class problem

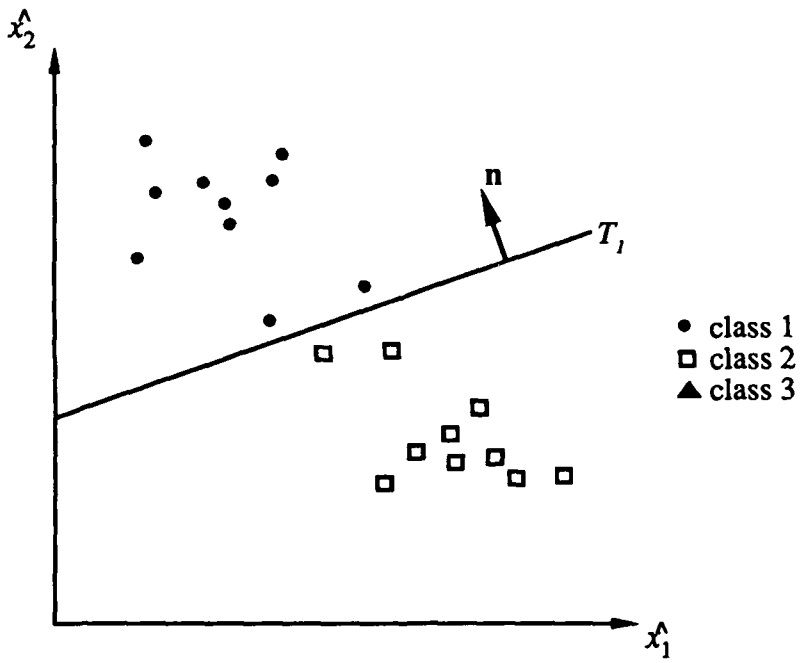


Figure 1.2 Classification with a linear decision boundary. Images above boundary T are taken to belong to class 1; those below are associated with class 2.

illustrated in Fig. 1.2, in which the $\hat{x}_i$-axis represents the intensity of pixel i in each image. The simplest approach to determining whether a particular image belongs to class 1 or to class 2 is to decide to which side of the line, T, its vector, $\mathbf{x}$, points. This can be accomplished by forming the inner product, C (= $\mathbf{n}\cdot\mathbf{x}$), between $\mathbf{x}$ and the normal, $\mathbf{n}$, to line T (in the general problem, T is a hyperplane). If the line T is described by the equation $\mathbf{n}\cdot\hat{\mathbf{x}} = k$, then the following decision rule applies:

Outcome	Decision
$C > k$	$\mathbf{x} \in$ class 1
$C < k$	$\mathbf{x} \in$ class 2
$C = k$	Don't know

In this method, T is known as the *decision boundary*, $\mathbf{n}$ is called the *discriminant vector*, and $C = D(\mathbf{x}) = \mathbf{n} \cdot \mathbf{x}$ is the *discriminant function.*

This approach can easily be extended to multiple-class problems by introducing

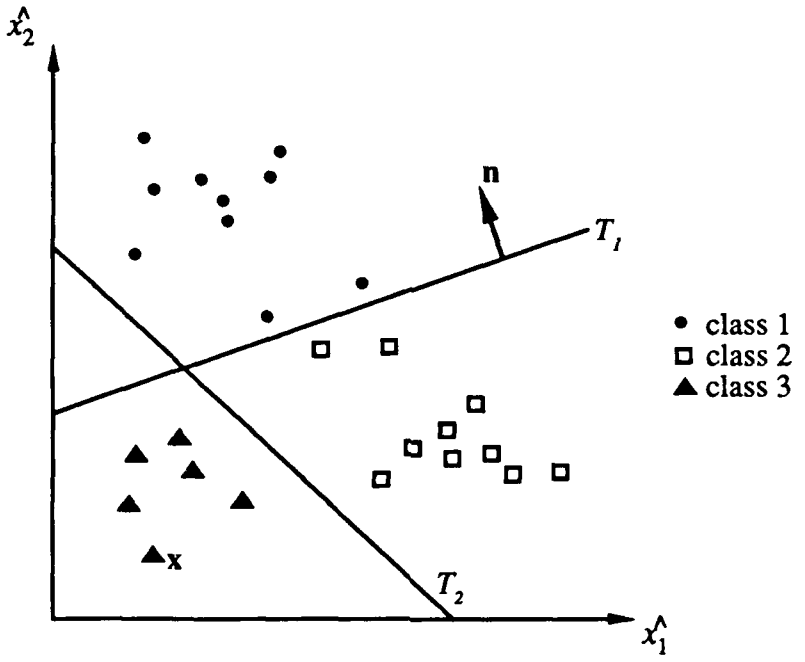


Figure 1.3 Multiple decision boundaries.
Class 3 elements are identified as those lying in a region of intersection of half-spaces defined by T_1 and T_2.

additional boundaries. In Fig. 1.3, we see that the input image vector, **x**, belongs to class 3 since it lies in the intersection of the region below T_1 and the region to the left of T_2. If K classes can be separated in this manner, then at most (K-1) boundaries are needed to distinguish them.

In the two-class example, the decision step in the classification was trivial; the inner product was simply compared to a threshold value. In the multiple-class problem, this step gains some significance. A logical AND operation must be carried out to determine in which intersection region the image lies. This operation is merely another inner product, simplified by the fact that the vectors to be operated on contain only binary components. If the number of boundaries is large, the problem of analyzing the inner product values may become significant if real-time operation is the goal.

The general approach outlined in this section has formed the backbone for many optical image classification methods described in the literature. The differences between various methods of this type lie in their approaches to choosing the boundary T, and in their schemes for calculating the inner product. In the following sections, representative methods for choosing the boundary will be described. Since the inner product is an effective way of extracting information about the position of a point in a multidimensional space, and since optics provides the means for its rapid computation, this thesis will focus on exploring the potential for inner-product based approaches, especially for real-time applications.

1.2 Approaches to discriminant vector synthesis

The heart of the problem of classifying images in their vector representation lies in the choice of the decision boundaries, or equivalently, of the discriminant vectors. The difficulty of the problem derives primarily from the large dimensionality of the spaces. A visual assessment of the two-dimensional example in Fig. 1.3, for example, immediately produces viable (though, perhaps, not optimal) solutions to the problem. The following discussions introduce a sample of existing approaches to the problem of discriminant vector synthesis and are intended to provide a background for the subsequent portions of this thesis.

1.2.1 The average filter

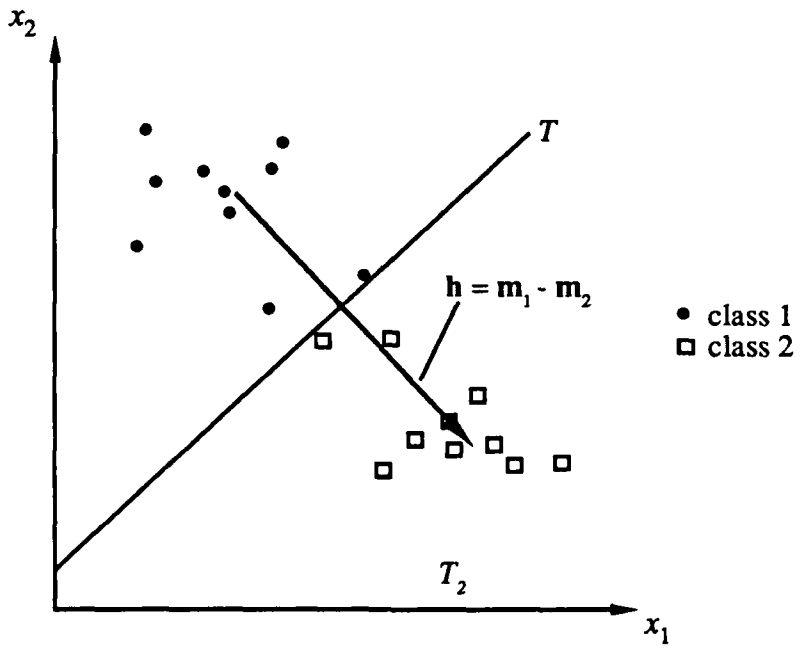


Figure 1.4 Difference of class means as discriminant vector.

The mean of a class is often loosely called the average filter. A simple, but often effective, choice of discriminant vector is the difference of the means of two class distributions. Figure 1.4 illustrates the geometrical significance of this approach. It is clear from the diagram that the idea behind the average filter is generally correct, but it is not optimal for general class distributions.

1.2.2 The data compression approach

Like many techniques, the method of Fukunaga and Koontz,[7] begins by assuming that the available examples of a class of signals are statistical realizations of an underlying distribution. This approach is broadly termed *statistical pattern recognition.* While the Fukunaga-Koontz transform was not originally intended for use in classifying images, it has been so utilized.[8]

In applying their method, we must think of elements of an image class as random perturbations within our multidimensional space about some mean image. We begin by writing the input image vector (now considered as a multidimensional random variable) in a basis vector expansion:

$$\mathbf{x} = \sum_{i=1}^{N} c_i \boldsymbol{\phi}_i \quad , \tag{1.7}$$

where the basis vectors ϕ_i are the eigenvectors of the autocorrelation matrix for the process, i.e.,

$$\mathbf{R}\boldsymbol{\phi}_i = \lambda_i \boldsymbol{\phi}_i \tag{1.8}$$

where

$$\mathbf{R} = E\{\mathbf{x}\mathbf{x}^T\} \quad . \tag{1.9}$$

In Eq. (1.9), $E\{...\}$ denotes an average over prototype images defining the image class. Projection of the random vector $\mathbf{x}$ onto this basis set produces a new vector for which the autocorrelation matrix is diagonalized. This transformation is known in various fields as the Karhunen-Loève transform,[9] principle components analysis, and factor analysis.

The Karhunen-Loève basis has the property that the mean-square error incurred in estimating $\mathbf{x}$ with only N' ($N' < N$) terms is minimized. The mean-square error is equal to the sum of the eigenvalues corresponding to the discarded components, i.e.,

$$<\varepsilon^2> = \sum_{i=N'+1}^{N} \lambda_i \quad . \tag{1.10}$$

The significance is that a few basis vectors can be chosen that are efficient in representing the class. They may not, however, be effective in discriminating the classes, especially classes characterized by similar autocorrelation matrices. The strategy of the Fukunaga-Koontz transform is to construct a single Karhunen-Loève basis set that contains information about both classes.

We begin by forming a pooled correlation matrix for two classes defined by

$$\mathbf{R}_t = P(\omega_1)\mathbf{R}_1 + P(\omega_2)\mathbf{R}_2 \quad , \tag{1.11}$$

where $P(\omega_j)$ represents the *a priori* probability for the jth class and $\mathbf{R}_j$ denotes the sample autocorrelation matrix for class j defined by

$$\mathbf{R}_j = \frac{1}{M_j} \sum_{k=1}^{M_j} \mathbf{x}_k^{(j)} \, \mathbf{x}_k^{(j)\,T} \quad , \tag{1.12}$$

where $\mathbf{x}_k^{(j)}$ is the kth prototype image for class j, and M_j is the number of prototypes for class j.

A transformation matrix **P** is determined that, when applied to an input image vector, yields a new vector for which the pooled correlation matrix is diagonalized according to

$$\mathbf{P}\mathbf{R}_t\mathbf{P}^T = \mathbf{I} \quad , \tag{1.13}$$

where **I** is the identity matrix. Applying the same transformation to the single-class autocorrelation matrices ($\mathbf{R}_1$ and $\mathbf{R}_2$), we find that the transformed matrices share a common set of eigenvectors (i.e., a single basis can be found for representing both classes according to the Karhunen-Loève formulation). In addition, the corresponding eigenvalues are related by

$$\lambda_i^{(1)} + \lambda_i^{(2)} = 1 \quad , \tag{1.14}$$

where $\lambda_i^{(j)}$ denotes the ith eigenvalue for class j. The net transformation is equivalent to simultaneous diagonalization of the pooled correlation matrix and one of the class autocorrelation matrices.

The componentwise relationship expressed in Eq. (1.14) indicates that the eigenvector directions that are most important for describing one class are least important for describing the other. Thus projection of the input image vector onto the basis defined by the common set of eigenvectors produces data that are likely to be useful for class discrimination.

1.2.3 Maximizing the average separation of projections

One approach to classification is to optimize a measure of the separation of the projections of the two classes onto the discriminant vector. The methods that follow

this approach differ primarily in their criteria for measuring the distance between the clusters of discriminant function values.

The most widely used methods in this category are the Foley-Sammon transform[10] and the Hotelling trace transform.[11] For the sake of brevity, only the former will be considered here.

In 1936, R.A. Fisher proposed a measure for the distance between clusters of taxonomic data.[12] This quantity, which has since become known as the Fisher ratio, is defined as

$$R = \frac{(m_1 - m_2)^2}{\sum_{j=1}^{2} \sum_{k=1}^{M_j} (y_k^{(j)} - m_j)^2} \quad , \tag{1.15}$$

where $y_k^{(j)}$ denotes the kth element of class j, m_j is the mean of the elements of class j, and M_j is the number of elements in class j. The numerator in Eq. (1.15) measures the distance between the clusters and the denominator measures their spread. Diffuse clusters or clusters that are close to one another are difficult to separate and are characterized by a small Fisher ratio. Widely spaced and compact clusters have large Fisher ratios.

If the $y_k^{(j)}$ data in Eq. (1.15) are values of the discriminant function in a classification problem, that is, if

$$y_k^{(j)} = \mathbf{x}_k^{(j)} \cdot \mathbf{h} \quad , \tag{1.16}$$

where $\mathbf{x}_k^{(j)}$ is the kth image vector from class j and $\mathbf{h}$ is the discriminant vector, then the Fisher ratio becomes

$$R(\mathbf{h}) = \frac{\left[\mathbf{h}^T(\mathbf{m}_1 - \mathbf{m}_2)\right]^2}{\sum_{j=1}^{2} \sum_{k=1}^{M_j} \left[\mathbf{h}^T(\mathbf{x}_k^{(j)} - \mathbf{m}_j)\right]^2} \quad , \tag{1.17}$$

or equivalently,

$$R(\mathbf{h}) = \frac{\mathbf{h}^T\mathbf{B}\mathbf{h}}{\mathbf{h}^T\mathbf{W}\mathbf{h}} \quad , \tag{1.18}$$

where $\mathbf{B}$ is the between-class scatter matrix of the image classes,

$$\mathbf{B} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \quad , \tag{1.19}$$

and $\mathbf{W}$ is the pooled the within-class scatter matrix,

$$\mathbf{W} = \sum_{j=1}^{2}\sum_{k=1}^{M_j}(\mathbf{x}_k^{(j)} - \mathbf{m}_j)(\mathbf{x}_k^{(j)} - \mathbf{m}_j)^T \quad . \tag{1.20}$$

The Fisher discriminant is the vector $\mathbf{h}$ that maximizes R, i.e., it is the discriminant vector that produces the largest average separation of the discriminant function values as measured by the Fisher ratio. It can be shown that a vector $\mathbf{h}$ that maximizes the Fisher ratio must satisfy the generalized eigenvector equation

$$\mathbf{B}\mathbf{h} = R\mathbf{W}\mathbf{h} \quad . \tag{1.21}$$

In the event that $\mathbf{W}$ is non-singular, Eq. (1.21) can be written in the form of an eigenvector equation

$$(\mathbf{W}^{-1}\mathbf{B})\mathbf{h} = R\mathbf{h} \quad . \tag{1.22}$$

In practice, $\mathbf{W}$ is usually singular and a pseudoinverse algorithm must be employed.

1.2.4 Mapping techniques

The approach of mapping techniques is to construct a system that produces specified values at its outputs when presented with images from the classes of interest. The most notable examples of this approach are the synthetic disciminant function[13] and the feed-foward neural network.[14,15]

1.2.4.1 The synthetic discriminant function

The synthetic discriminant function approach to classificatic ` to compute K discriminant vectors (corresponding to K image classes) that have the property that

$$\mathbf{h}_j \cdot \mathbf{x}_k^{(j')} = \delta_{jj'} \quad , \tag{1.23}$$

where $\mathbf{h}_j$ is the discriminant vector corresponding to class j, $\mathbf{x}_k^{(j')}$ is the kth prototype image for class j', and $\delta_{jj'}$ is the Kronecker δ-function. In other words, the goal of each synthetic discriminant vector is to give a unit response to an image belonging to the class to which it corresponds, and to give a zero response otherwise. Discriminant vectors that obey Eq. (1.23) are easily obtained if they are assumed to be linear combinations of the training images, i.e., if it is assumed that

$$\mathbf{h}_j = \sum_{j'=1}^{K} \sum_{k=1}^{M_{j'}} a_k^{(j,j')} \mathbf{x}_k^{(j')} \quad , \tag{1.24}$$

where the $a_k^{(j,j')}$ are real numbers. In this case, the problem of determining the discriminant vector reduces to that of computing the $a_k^{(j,j')}$ coefficients.

If the training images are arranged into the columns of a matrix $\mathbf{W}$ given by

$$\mathbf{W} = \left[\mathbf{x}_1^{(1)} \vdots \mathbf{x}_2^{(1)} \vdots \cdots \vdots \mathbf{x}_{M_1}^{(1)} \vdots \mathbf{x}_1^{(2)} \vdots \mathbf{x}_2^{(2)} \vdots \cdots \vdots \mathbf{x}_{M_2}^{(2)} \vdots \cdots \vdots \mathbf{x}_1^{(K)} \vdots \mathbf{x}_2^{(K)} \vdots \cdots \vdots \mathbf{x}_{M_K}^{(K)}\right], \tag{1.25}$$

then the linear combination condition can be written as

$$\mathbf{h}_j = \mathbf{W}\mathbf{b}_j \quad , \tag{1.26}$$

where $\mathbf{b}_j = \left[a_1^{(j,1)}, a_2^{(j,1)}, \ldots, a_{M_1}^{(j,1)}, a_1^{(j,2)}, a_2^{(j,2)}, \ldots, a_{M_2}^{(j,2)}, \ldots, a_1^{(j,K)}, a_2^{(j,K)}, \ldots, a_{M_K}^{(j,K)}\right]^T$.

Likewise the output condition (Eq. (1.23)) can be written as

$$\mathbf{W}^T\mathbf{h}_j = \mathbf{u}_j \quad , \tag{1.27}$$

where $\mathbf{u}_j$ is a vector composed of zeros in each component except the jth which contains a one. Combining Eqs. (1.26) and (1.27), we obtain

$$\mathbf{W}^T\mathbf{W}\mathbf{b}_j = \mathbf{u}_j \quad . \tag{1.28}$$

If the matrix $(\mathbf{W}^T\mathbf{W})$ is non-singular, then the solutions for the coefficients in the linear combination of Eq. (1.24) are given by

$$\mathbf{b}_j = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{u}_j \quad . \tag{1.29}$$

1.2.4.2 The feed-forward neural network

The neural network approach to pattern classification[14,15] is somewhat different from the other methods outlined in this chapter. It is founded on the same building block (the inner product followed by a threshold operation) but, unlike the other methods, it uses multiple layers of decisions. In the first level of operation of the feed-forward network, multiple discriminant vectors are applied to the input image. The resulting inner products are passed through a soft threshold function and the results form the input for another layer of discriminant vectors. The results of the next step are soft-thresholded and can form the input to still another layer, and so on. For comparison, the standard pattern recognition approaches described in previous sections can be described as one-layer networks.

The standard approach to generating the discriminant vectors in the feed-forward architecture is an iterative scheme known as the back-propagation algorithm. In this approach, as in the synthetic discriminant function method, the outputs are used to encode the classification decision. In this technique, however, a gradient descent algorithm is used to optimize the output, i.e., to minimize the mean squared error between some specified desired output and the actual output for each prototype image. To "train" the network, prototype images are sequentially presented at the inputs, and

the components of the discriminant vectors, known as interconnect strengths, are adjusted according to a well-defined prescription to optimize the mean-square-error criterion.

1.3 Overview of the thesis

In Chapter 2 a new approach to discriminant vector synthesis for image classification is described. The method applies convex analysis to the problem of boundary selection and, thus, deviates from the statistical approaches that have characterized the optical pattern recognition literature in the past. Basic terms from set theory are defined in Section 2.1 to introduce the discussion. In Section 2.2 the basic properties of the proposed discriminant vector are derived. In particular, the discriminant vector is shown to guarantee separation of the image classes if they are, indeed, linearly separable. Further it is shown to maximize the minimum separation of the discriminant function values for two classes. In Section 2.3 the problem of computing the discriminant vector is defined. It is shown to be the solution of a constrained optimization problem that scales in complexity with the number of prototype images used to define the classes. This represents an advantage over many other discriminant-vector approaches that involve computations that scale with the number of pixels in the image. In Section 2.4, quadratic programming, the particular type of constrained optimization involved in the proposed method is reviewed. In Section 2.5 the effects of the dimensionality of the space defined by the pixels are considered and approaches to discrimination of multiple classes are introduced. Section 2.6 contains a description of experiments that demonstrate the proposed method for a two-class and an eight-class classification problem. The results obtained in both cases are extremely promising.

Chapter 3 contains a development of approaches to quantum-limited image classification, a problem that has not previously been considered in any detail. Quantum-limited images arise in night vision, low-dose electron microscopy, and radiological imaging. The ability to identify severely degraded images would have great significance for these applications. In addition, certain aspects of quantum-limited images and detection systems suggest that there might be advantages gained from their use in ordinary high-light-level situations. In Section 3.1, the photon-counting detector used in the research is described. In Section 3.2 the statistical properties of spatially quantized quantum-limited images are described. In Section 3.3 the inner product between a quantum-limited image and a discriminant vector is considered and an approach to its computation is outlined. In Section 3.4 the statistical properties of the quantum-limited inner product are discussed. Sections 3.3 and 3.4 demonstrate that any linear discriminant function can be implemented using a quantum-limited imaging system, however new approaches based on statistical decision theory, are presented in Section 3.6 with an introductory review of decision theory given in Section 3.5. In Section 3.6.1 the solutions developed in Section 3.6 are demonstrated experimentally using a commercial, photon-counting detector, packaged as a camera by the author and colleagues. The results demonstrate that a very small number of photoevents detected in the image plane can be sufficient to make reliable classification decisions. Further, for the images tested, the statistical decision theory method developed in Section 3.6 is shown to permit classifications with less photons than the Fukunaga-Koontz and average filter approaches. In Section 3.7 the statistical decision theory approach is combined with an invariant-filtering technique from the literature to permit classification

of rotated images. This constitutes the first attempt to combine classification discrimant vectors with filters for geometrically-invariant recognition. In Section 3.7.1 the proposed method is demonstrated in experiments involving quantum-limited images. It is shown that the addition of rotation-invariance demands an increase in the number of detected photoevents for reliable classification, but the number is still very small. Chapter 3 concludes, in Section 3.8, with a discussion of how the low-light-level solutions derived in Section 3.7 might be applied at high light levels.

Chapter 4 contains a summary of the work and a brief discussion of future directions in the area of optical image classification research.

Much of the work described in Chapter 3 has been published and/or reported in conference proceedings (see Refs. 16-19). The work contained in Chapter 2 is relatively new and has not yet been submitted for publication.

References for Chapter 1

1. Q. Tian, Y. Fainman, Z. H. Gu, and Sing. H. Lee, "Comparison of statistical pattern-recognition algorithms for hybrid processing. I. Linear-mapping algorithms," J. Opt. Soc. Am. A, **5**, 10, 1655-1669 (1988).
2. Q. Tian, Y. Fainman, and Sing. H. Lee, "Comparison of statistical pattern-recognition algorithms for hybrid processing. I. Eigenvector-based algorithms," J. Opt. Soc. Am. A, **5**, 10, 1670-1682 (1988).
3. N. Nandhakumar and J. K. Aggarwal, "The artificial intelligence approach to pattern recognition - a perspective and overview," Patt. Recog. **18**, 383-389 (1985).
4. D. H. Ballard and C. M. Brown, *Computer Vision*, (Prentice-Hall, Englewood Cliffs, NJ).
5. Robert Fiete, Ph.D. Thesis, University of Arizona.
6. A. Vander Lugt, IEEE Trans. Inf. Theory **IT-10**, 139 (1964).
7. K. Fukunaga and W. L. G. Koontz, "Application of the Karhunen-Loève expansion to feature selection and ordering," IEEE Trans. Comput. **C-19**, 311-318 (1970).
8. J. R. Leger and S. H. Lee, "Image classification by an optical implementation of the Fukunaga-Koontz transform," J. Opt. Soc. Am. **72**, 556-564 (1982).
9. N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing* (Springer-Verlag, New York, 1975).
10. D. H. Foley and J. W. Sammon, Jr., "An optimal set of discriminant vectors," IEEE Trans. Comput. **C-24**, 281-289 (1975).

11. K. Fukanaga, *Introduction to Statistical Pattern Recognition* (Academic Press, New York, 1972), p. 260 *ff*.

12. R. A. Fisher, "The use of multiple measurements in taxonomic problems," Ann. Eugen. **7**, 179-188 (1936).

13. D. Casasent, "Unified synthetic discriminant function computational formulation," Appl. Opt. **23**, 1620-1627 (1984).

14. L. Saaf and G. Michael Morris, "Filter synthesis using neural networks," Proc. SPIE **1134**, 12-16 (1989).

15. D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1 (MIT Press, Cambridge, 1986).

16. Miles N. Wernick and G. Michael Morris, "Image Classification at Low Light Levels," J. Opt. Soc. Am. A, **3**, 2179 (1986).

17. Miles N. Wernick and G. Michael Morris, "Maximum-Likelihood Image Classification," Proc. SPIE, **938** (1988).

18. Miles N. Wernick, Thomas A. Isberg, and G. Michael Morris, "Rotation-invariant image classification," J. Opt. Soc. Am. A, **3**, P86 (1986).

19. Miles N. Wernick and G. Michael Morris, "Image Classification at Low Light Levels," Technical Digest, O.S.A. Topical Meeting on Quantum-Limited Imaging and Image Processing, Honolulu, 106 (1986).

Chapter 2:
Pattern classification by separation of convex hulls

As outlined in Chapter 1, many traditional pattern classification methods are based on statistical averages over the image classes. One drawback of these approaches, for problems in which the classes are not truly statistical in nature, is that they tend to give little weight to outlier images.

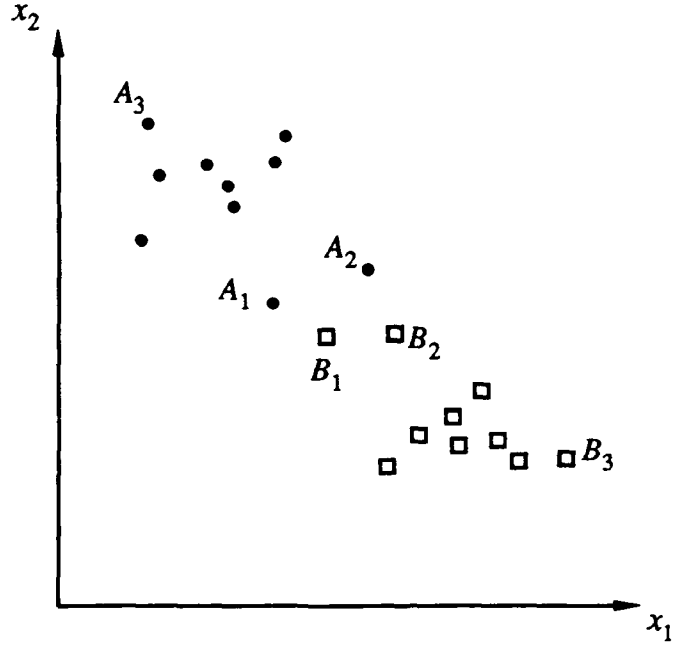


Figure 2.1 Hypothetical class distributions. Images A_1, A_2, B_1, and B_2 may be misclassified by a method that relies on statistical averages.

The scatter plot in Fig. 2.1 represents a hypothetical distribution of two image classes that is exaggerated to demonstrate this point. In Figure 2.1, as in Chapter 1, x_i

represents the intensity of pixel i. A method that uses statistical averages to choose the discriminant vector will give little emphasis to images A_1, A_2, B_1, and B_2, although these are the points that are most likely to be misclassified. Images as far distant as A_3 and B_3 are likely to be correctly classified by any reasonable choice of discriminant vector and do not deserve the emphasis they generally receive. In contradistinction to the approaches described in Chapter 1, the method presented in this chapter is designed to emphasize the elements of each class that are most difficult to classify. The effect achieved by the proposed approach is to provide a discriminant vector that produces no classification errors if the two classes are strictly linearly separable and to indicate that no such solution exists if they are not.

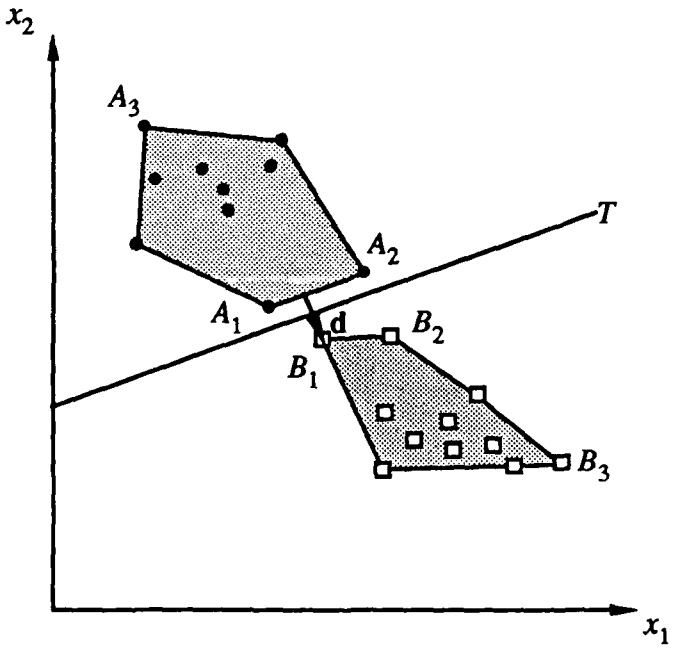


Figure 2.2 Construction of separating hyperplane. The line segment connecting the points of closest approach of the regions is normal to the boundary T.

To illustrate the approach in an intuitive way, let us consider again the hypothetical class distributions plotted in Fig. 2.1. If we were asked to choose, by inspection, a decision boundary for the two classes, we might first associate with each cluster of points a bounded region, as shown in Fig. 2.2. A natural choice for the boundary between classes might then be the perpendicular bisector of the line segment that connects the points of closest approach of the two regions. The vector direction **d** along which our line segment lies is a discriminant vector for the two-class problem and the bisector T (a hyperplane in actual problems) is an effective decision boundary.

In its precise form, this intuitive method to boundary selection is a well-known construction from the theory of convex analysis.[1,2] In Section 2.2, we will show that it is a solution to the image classification problem that is optimum in the sense of providing the space in which the minimum separation of the two classes is maximized.

In 1965, Rosen[3] proposed a different formulation of the problem for general pattern classification problems that produces the same discriminant vector as that proposed. Rosen's formulation, however, leads to an intractable computation for the number of dimensions typical of image classification problems. A brief comparison of the proposed solution with Rosen's is given in Appendix A.

In the following section, basic definitions from convex analysis are reviewed to introduce the discussion. In Section 2.2, the separation and optimality properties of the proposed method are developed. In Section 2.3, the proposed discriminant vector is shown to be the solution of a particular quadratic program; a brief overview of quadratic programming is given in Section 2.4. Section 2.5 addresses the effects of the

dimensionality of the pixel space. In Section 2.6, experimental results demonstrate application of the method to actual images.

2.1 Convex sets and linear separability

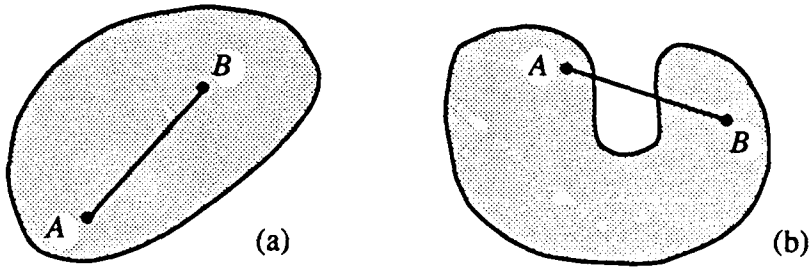


Figure 2.3 Convex (a) and nonconvex (b) regions.

We begin with a brief and informal review of relevant terminology from set theory that will be used in subsequent sections. In a real Euclidean space R^n, a set is said to be *convex* if the line segment connecting any two points in the set also lies entirely within the set. The set shown in Fig. 2.3(a) is convex; the set shown in Fig. 2.3(b) is not. The *convex hull* of a set S, denoted by conv S, is the smallest convex set containing S. In R^2, conv S can be thought of as the region enclosed by a "rubber band" stretched around S as illustrated in Fig. 2.4.

The sets considered in the proposed method consist of a finite collection of points representing the training images. The convex hull of a set consisting of M points, not lying in one (M-2)-dimensional plane, is a polyhedral region known as an (M-1)-*dimensional simplex*. A two-dimensional simplex is a triangle; a three-dimensional simplex is a tetrahedron. The convex hull of a finite point set is also

known as its *convex polytope*. The convex polytope of a set $A = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k)$ is the set of points $\mathbf{x}$ such that

$$\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{a}_i \quad , \tag{2.1}$$

where

$$\lambda_i \geq 0 \quad (i = 1, 2, \ldots, k) \tag{2.2}$$

and

$$\sum_{i=1}^{k} \lambda_i = 1 \quad . \tag{2.3}$$

The λ_i are known as the *barycentric coordinates* of $\mathbf{x}$.

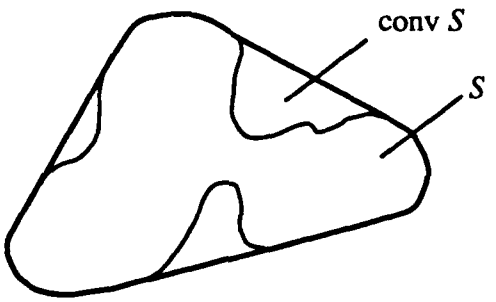


Figure 2.4. A set S and its convex hull, conv S .

Consistent with the discussions of Chapter 1, the aim of the proposed method is to separate the two image classes with a hyperplane. If S_1 and S_2 are sets in R^n we will define them as *strictly linearly separable* if there exists a hyperplane T such that S_1 lies in one of the *open* half-spaces defined by T, and S_2 lies in the other. *Weak separability* permits the hyperplane to contain a portion of the boundary of one or both sets.

2.2 Properties of the convex-hull discriminant vector

In this section, the fundamental properties of the proposed method are demonstrated. First, supposing point sets representing two image classes and given the convex hulls of these point sets, it will be shown that the ability of a certain hyperplane to separate the convex hulls is a necessary and sufficient condition for that hyperplane to separate the point sets. Then it will be shown that if, in fact, the convex hulls are separable, the construction described in the introduction of this chapter (see Fig. 2.2) is guaranteed to find such a hyperplane (i.e., one that separates the convex hulls and, hence, the classes). Finally, we will prove that the solution provided by the proposed method not only separates the classes, but is optimum in the following sense. If we form the inner product between each prototype (or training) image and the discriminant vector, we will obtain two sets of values (one for each class). The proposed discriminant vector is the one for which the minimum separation of these two sets of values is maximized, i.e., it produces the largest gap between the two groups.

We begin by presenting a theorem that demonstrates that separation of the point sets is equivalent to separation of their convex hulls, hence, that solution of the convex-hull separation problem implies solution of the classification problem.

Theorem. Let $A = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k)$ and $B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_l)$ be finite collections of points in R^n. The strict linear separability of the convex hulls of A and B (by a hyperplane $\mathbf{p} \cdot \mathbf{x} = \alpha$) is a necessary and sufficient condition for strict linear separability of A and B (by the same hyperplane).

Proof. (Necessity) If A and B are strictly linearly separable, then by definition there exists a hyperplane $\mathbf{p}\cdot\mathbf{x}=\alpha$ such that

$$\begin{aligned}\mathbf{p}\cdot\mathbf{a}_i &> \alpha \quad ;i=1,2,\ldots,k\\ \mathbf{p}\cdot\mathbf{b}_j &< \alpha \quad ;j=1,2,\ldots,l \quad .\end{aligned} \tag{2.4}$$

If $\mathbf{y}^{(1)} \in \text{conv } A$, then $\mathbf{y}^{(1)}$ can be written as

$$\mathbf{y}^{(1)}=\sum_{i=1}^{k}\lambda_i\mathbf{a}_i \quad , \tag{2.5}$$

where

$$\lambda_i \geq 0 \quad (i=1,\ 2,\ \ldots\ ,\ k) \tag{2.6}$$

and

$$\sum_{i=1}^{k}\lambda_i=1 \quad . \tag{2.7}$$

Multiplying Eq. (2.5) by $\mathbf{p}$, we obtain

$$\mathbf{p}\cdot\mathbf{y}^{(1)}=\sum_{i=1}^{k}\lambda_i(\mathbf{p}\cdot\mathbf{a}_i) \quad . \tag{2.8}$$

Using Eq. (2.4), it follows that

$$\mathbf{p}\cdot\mathbf{y}^{(1)}>\sum_{i=1}^{k}\lambda_i\alpha \quad . \tag{2.9}$$

Factoring α from the summation and using Eq. (2.7), we obtain

$$\mathbf{p}\cdot\mathbf{y}^{(1)}>\alpha \quad , \tag{2.10}$$

for all $\mathbf{y}^{(1)} \in \text{conv } A$. By a similar argument it can be shown that for any point $\mathbf{y}^{(2)} \in \text{conv } B$,

$$\mathbf{p}\cdot\mathbf{y}^{(2)}<\alpha \quad . \tag{2.11}$$

Therefore, if A and B are strictly separable (by the hyperplane $\mathbf{p}\cdot\mathbf{x}=\alpha$) then conv A and conv B are also strictly separable (by the same hyperplane).

(Sufficiency) Suppose that the hyperplane $\mathbf{p}\cdot\mathbf{x} = \alpha$ strictly separates conv A from conv B. Since $A \subset \text{conv } A$ and $B \subset \text{conv } B$, $\mathbf{p}\cdot\mathbf{x} = \alpha$ also separates A from B.

The proof of the following theorem demonstrates that the proposed method for separation of the convex hulls guarantees a discriminant vector solution provided that one exists. The proof is essentially identical to one given in Ref. 4 for the Minkowski separation theorem.

Theorem. Let $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ be any two points drawn from conv A and conv B, respectively, and let $\bar{\mathbf{y}}^{(1)}$ and $\bar{\mathbf{y}}^{(2)}$ be the points that minimize the distance function $\|\mathbf{y}^{(1)}-\mathbf{y}^{(2)}\|$. Let $\mathbf{p} = \bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}$, i.e., let $\mathbf{p}$ lie along the line segment connecting points $\bar{\mathbf{y}}^{(1)}$ and $\bar{\mathbf{y}}^{(2)}$. Then the vector $\mathbf{p}$ is the norm of a hyperplane that separates the convex hulls, provided that the convex hulls are non-overlapping.

Proof. If the convex hulls do not overlap, i.e., if for all $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, $\mathbf{y}^{(1)} \neq \mathbf{y}^{(2)}$, then

$$\|\mathbf{p}\|^2 > 0 \quad , \tag{2.12}$$

or,

$$\mathbf{p}\cdot\mathbf{p} > 0 \quad . \tag{2.13}$$

Using the definition of $\mathbf{p}$, we find

$$\mathbf{p}\cdot[\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}] > 0 \quad , \tag{2.14}$$

or,

$$\mathbf{p}\cdot\bar{\mathbf{y}}^{(1)} > \mathbf{p}\cdot\bar{\mathbf{y}}^{(2)} \quad . \tag{2.15}$$

What remains to be shown is that $\mathbf{p}\cdot\mathbf{y}^{(1)} \geq \mathbf{p}\cdot\bar{\mathbf{y}}^{(1)}$ and $\mathbf{p}\cdot\bar{\mathbf{y}}^{(2)} \geq \mathbf{p}\cdot\mathbf{y}^{(2)}$ for all $\mathbf{y}^{(1)} \in$ conv A and $\mathbf{y}^{(2)} \in$ conv B.

We can construct a parametric representation for points in each convex hull in the following way. Define $\mathbf{y}_\lambda^{(2)} \in$ conv B as

$$\mathbf{y}_\lambda^{(2)} = (1-\lambda)\bar{\mathbf{y}}^{(2)} + \lambda \mathbf{y}^{(2)} \quad . \tag{2.16}$$

Then

$$\left\|\bar{\mathbf{y}}^{(1)} - \mathbf{y}_\lambda^{(2)}\right\|^2 = [\lambda(\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)}) + (1-\lambda)(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)})] \cdot [\lambda(\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)}) + (1-\lambda)(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)})] \quad , \tag{2.17}$$

or, expanding the inner product,

$$\left\|\bar{\mathbf{y}}^{(1)} - \mathbf{y}_\lambda^{(2)}\right\|^2 = (1-\lambda)^2 \left\|\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right\|^2 + 2\lambda(1-\lambda) \times [(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}) \cdot (\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)})] + \lambda^2 \left\|\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)}\right\|^2 . \tag{2.18}$$

Differentiating with respect to λ and evaluating at $\lambda = 0$ gives

$$\frac{\partial}{\partial \lambda} \left\|\bar{\mathbf{y}}^{(1)} - \mathbf{y}_\lambda^{(2)}\right\|^2 \Bigg|_{\lambda=0} = -2\left\|\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right\|^2 + 2(\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)})(\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)}) \tag{2.19}$$

$$= -2\mathbf{p} \cdot (\mathbf{y}^{(2)} - \bar{\mathbf{y}}^{(2)}) \quad . \tag{2.20}$$

Since $\bar{\mathbf{y}}^{(2)}$ is the point that minimizes $\left\|\bar{\mathbf{y}}^{(1)} - \mathbf{y}^{(2)}\right\|^2$ on conv B, the derivative must be nonnegative, i.e.,

$$-2\mathbf{p} \cdot (\mathbf{y}^{(2)} - \bar{\mathbf{y}}^{(2)}) \geq 0 \quad , \tag{2.21}$$

or,

$$\mathbf{p} \cdot \bar{\mathbf{y}}^{(2)} \geq \mathbf{p} \cdot \mathbf{y}^{(2)} \quad . \tag{2.22}$$

A similar argument for conv A shows that

$$\mathbf{p} \cdot \mathbf{y}^{(1)} \geq \mathbf{p} \cdot \bar{\mathbf{y}}^{(1)} \quad . \tag{2.23}$$

Combined with inequality (2.15), inequalities (2.22) and (2.23) imply, by definition, the linear separability of conv A and conv B by any of the planes $\mathbf{p} \cdot \mathbf{x} = \alpha$ where α lies in the interval $\mathbf{p} \cdot \bar{\mathbf{y}}^{(2)} < \alpha < \mathbf{p} \cdot \bar{\mathbf{y}}^{(1)}$.

We conclude this section by showing that the proposed solution is the one that maximizes the minimum separation of the projections of the two classes onto the discriminant vector. In other words, the proposed discriminant vector is the value of $\mathbf{p}$ that solves the following optimization problem:

Maximize δ with respect to $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$, and $\mathbf{p}$

subject to the conditions

$$\mathbf{p}\cdot\mathbf{y}^{(1)} - \mathbf{p}\cdot\mathbf{y}^{(2)} \geq \delta \quad , \tag{2.24a}$$

$$\|\mathbf{p}\| = 1 \quad , \tag{2.24b}$$

$$\text{for all } \mathbf{y}^{(1)} \in \text{conv}\,A \text{ and } \mathbf{y}^{(2)} \in \text{conv}\,B \quad . \tag{2.24c}$$

Since δ has no specific dependence on the choice variables, we need only find the greatest lower bound on the left-hand-side of inequality (2.24a). We can rewrite the left-hand-side of (2.24a) as follows:

$$\mathbf{p}\cdot\mathbf{y}^{(1)} - \mathbf{p}\cdot\mathbf{y}^{(2)} = \mathbf{p}\cdot[\mathbf{y}^{(1)} - \mathbf{y}^{(2)}] \tag{2.25}$$

$$= \|\mathbf{p}\|\,\|\mathbf{y}^{(1)} - \mathbf{y}^{(2)}\| \cos[\mathbf{p}, \mathbf{y}^{(1)} - \mathbf{y}^{(2)}] \quad . \tag{2.26}$$

Applying constraints (2.24b) and (2.24c), we find that δ is constrained by

$$\delta \leq \|\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\| \cos[\mathbf{p}, \bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}] \quad , \tag{2.27}$$

where, as before, $\bar{\mathbf{y}}^{(1)}$and $\bar{\mathbf{y}}^{(2)}$ are the points that minimize the distance function $\|\mathbf{y}^{(1)}\text{-}\mathbf{y}^{(2)}\|$. The maximum value of δ that satisfies the constraints is obtained by choosing $\mathbf{p}$ such that $\cos[\mathbf{p}, \bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}]$is maximized. This, of course, occurs for any $\mathbf{p}$ that lies along the direction of $\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}$ but applying (2.24b) we find that the optimum value, $\mathbf{p}_0$, is given by

$$\mathbf{p}_0 = \frac{\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}}{\left\|\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right\|} \quad , \tag{2.28}$$

and the maximum separation, δ, is given by

$$\delta = \left\|\bar{\mathbf{y}}^{(1)} - \bar{\mathbf{y}}^{(2)}\right\| \quad . \tag{2.29}$$

The optimum vector $\mathbf{p}_0$ is the proposed discriminant vector and is simply the unit vector along the line segment connecting the points of closest approach of the convex hulls for two classes.

2.3 Computation of the discriminant vector

The discriminant vector that we wish to compute lies along the line segment joining the points of closest approach of the convex hulls of two point sets representing training or prototype images for two image classes. If the jth image class contains M_j images, $\mathbf{x}_1, \ldots, \mathbf{x}_{Mj}$, then the convex hull for class j is the set of points $\mathbf{y}^{(j)}$ such that

$$\mathbf{y}^{(j)} = \sum_{k=1}^{M_j} \lambda_k^{(j)} \mathbf{x}_k^{(j)} \quad , \tag{2.30}$$

where the $\lambda_k^{(j)}$ are constrained to be non-negative and must sum to unity for each class, i.e.,

$$\sum_{k=1}^{M_j} \lambda_k^{(j)} = 1 \qquad (j = 1,2) \quad . \tag{2.31}$$

We now consider the problem of finding the points of closest approach of the two convex hulls. The squared Euclidean distance between points $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ drawn from the convex hulls of classes 1 and 2, respectively, is defined in the usual way as

$$d^2 = \sum_i [y_i^{(1)} - y_i^{(2)}]^2 \quad , \tag{2.32}$$

where i indexes the components of the vectors (here, the pixels of the images). Substituting from Eq. (2.30) into Eq. (2.32) we obtain

$$d^2 = \sum_i \left[\sum_{k=1}^{M_1} \lambda_k^{(1)} x_{k,i}^{(1)} - \sum_{k=1}^{M_2} \lambda_k^{(2)} x_{k,i}^{(2)} \right]^2 \quad , \tag{2.33}$$

where, as before, $x_{k,i}^{(j)}$ denotes the value of the ith pixel in the kth image of class j. Expanding the square in Eq. (2.33) and interchanging the order of summation yields

$$d^2 = \sum_{j=1}^{2} \sum_{k=1}^{M_j} \sum_{k'=1}^{M_j} \lambda_k^{(j)} \lambda_{k'}^{(j)} \sum_i x_{k,i}^{(j)} x_{k',i}^{(j)} - 2 \sum_{k=1}^{M_1} \sum_{k'=1}^{M_2} \lambda_k^{(1)} \lambda_{k'}^{(2)} \sum_i x_{k,i}^{(1)} x_{k',i}^{(2)} \quad . \tag{2.34}$$

This quadratic form can be written in matrix notation as

$$d^2 = \boldsymbol{\lambda}^T \mathbf{C} \boldsymbol{\lambda} \tag{2.35}$$

where the vector $\boldsymbol{\lambda}$ is defined as

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda_1^{(1)} \\ \vdots \\ \lambda_{M_1}^{(1)} \\ \lambda_1^{(2)} \\ \vdots \\ \lambda_{M_2}^{(2)} \end{bmatrix} , \tag{2.36}$$

and the symmetric matrix $\mathbf{C}$ is of the form

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} , \tag{2.37}$$

where the diagonal blocks, $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$, are of the form

$$\mathbf{C}_{jj} = \begin{bmatrix} \mathbf{x}_1^{(j)} \cdot \mathbf{x}_1^{(j)} & & \\ \vdots & \ddots & \\ \mathbf{x}_{M_j}^{(j)} \cdot \mathbf{x}_1^{(j)} & \cdots & \mathbf{x}_{M_j}^{(j)} \cdot \mathbf{x}_{M_j}^{(j)} \end{bmatrix} , \tag{2.38}$$

and the off-diagonal blocks, $\mathbf{C}_{21}$ and $\mathbf{C}_{12}$, are given by

$$\mathbf{C}_{21}{}^T = \mathbf{C}_{12} = \begin{bmatrix} -\mathbf{x}_1^{(1)} \cdot \mathbf{x}_1^{(2)} & & \\ \vdots & \ddots & \\ -\mathbf{x}_{M_1}^{(1)} \cdot \mathbf{x}_1^{(2)} & \cdots & -\mathbf{x}_{M_1}^{(1)} \cdot \mathbf{x}_{M_2}^{(2)} \end{bmatrix} . \tag{2.39}$$

In Eqs. (2.38) and (2.39), the following shorthand notation applies:

$$\mathbf{x}_k^{(j)} \cdot \mathbf{x}_{k'}^{(j')} = \sum_i x_{k,i}^{(j)} x_{k',i}^{(j')} . \tag{2.40}$$

We wish to find the points drawn from the two convex hulls that minimize the distance function d^2. We have expressed these points in terms of the λ-coefficients and have therefore reduced the problem to one of constrained optimization:

Minimize with respect to $\boldsymbol{\lambda}$ the distance function

$$d^2 = \boldsymbol{\lambda}^T \mathbf{C} \boldsymbol{\lambda} \tag{2.41}$$

subject to

$$\sum_{k=1}^{M_j} \lambda_k^{(j)} = 1 \qquad (j = 1,2) \tag{2.42}$$

and

$$\lambda_k^{(j)} \geq 0 \qquad (j = 1,2; k = 1,\ldots,M_j) \quad . \tag{2.43}$$

Any problem of this type, in which the optimization of a quadratic function is subject to linear and non-negativity constraints, is known as a *quadratic program*. The theory of quadratic programming is well developed and several techniques for the solution of such problems have been devised.[5] An overview of quadratic programming is provided in the following section.

2.4 Quadratic programming

The general convex quadratic programming problem is one of constrained optimization having the following form:

Maximize with respect to the vector $\mathbf{x}$ the function

$$F(\mathbf{x}) = \mathbf{p}^T\mathbf{x} - \tfrac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} \tag{2.44}$$

subject to the linear equality constraints

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.45}$$

and the non-negativity constraints

$$x_i \geq 0 \quad . \tag{2.46}$$

In expressions (2.44)-(2.46), $\mathbf{x}$ is an n-dimensional column vector $(x_1,\ldots, x_n)$, $\mathbf{p}$ is an n-dimensional column vector, $\mathbf{b}$ is an m-dimensional column vector, $\mathbf{A}$ is an $m \times n$ matrix ($m < n$), and $\mathbf{C}$ is an $n \times n$ symmetric positive semi-definite matrix.

The problem stated in other words is to find the point on a hyperplane in the non-negative orthant of R^n at which the function F attains a relative maximum. Each constraint forces the solution to lie in a hyperplane of smaller dimension. For example, if the problem contains two equality constraints in the form of three-dimensional planes, the solution is constrained to lie along the portion of the line representing their intersection that is contained in the non-negative orthant. If the planes do not intersect, the constraints are inconsistent and there is, of course, no solution.

Minimization of a function $f(\mathbf{x})$ can be placed in the above context by considering it as a problem of maximizing $-f(\mathbf{x})$. Likewise inequality constraints can be converted to equality constraints by introducing a vector of "slack variables", $\mathbf{y}$. For

example, the constraint set $\mathbf{Ax} \leq \mathbf{b}$ is equivalent to two sets of constraints: $\mathbf{Ax} + \mathbf{y} = \mathbf{b}$ and $y_i \geq 0$. The slack variables in $\mathbf{y}$ can be considered as additional x-variables.

The solution method for the quadratic program begins with the definition of the Lagrangean. The Lagrangean expression for the above problem[6] is

$$\phi(\mathbf{v},\mathbf{x}) = \mathbf{p}'\mathbf{x} - \tfrac{1}{2}\mathbf{x}'\mathbf{C}\mathbf{x} - \mathbf{v}'(\mathbf{Ax} - \mathbf{b}) \quad , \tag{2.47}$$

where $\mathbf{v}$ is a vector of m Lagrange multipliers. The gradient of the Lagrangean with respect to $\mathbf{x}$, denoted by $\mathbf{u}$, is given by

$$\mathbf{u} = \mathbf{p} - \mathbf{Cx} - \mathbf{A}'\mathbf{v} \quad . \tag{2.48}$$

The conditions for optimality of a solution $\mathbf{x}$, known as the Kuhn-Tucker conditions, are

$$\mathbf{u} = \mathbf{p} - \mathbf{Cx} - \mathbf{A}'\mathbf{v} \quad , \tag{2.49}$$

$$\mathbf{Ax} = \mathbf{b} \quad , \tag{2.50}$$

$$x_i \geq 0 \quad , \tag{2.51}$$

$$\mathbf{u}'\mathbf{x} = 0 \quad , \tag{2.52}$$

$$u_i \leq 0 \quad . \tag{2.53}$$

Since both the objective function and the non-negativity constraint are differentiable and convex, these conditions are both necessary and sufficient.[7]

The first Kuhn-Tucker condition is simply the definition of $\mathbf{u}$. The next two are merely restatements of the constraints. The last two conditions summarize three possible situations that lead to a solution of the problem. These are best illustrated by considering a single-variable problem having non-negativity constraints but no linear constraints (in this case, $\mathbf{u}$ is simply the gradient of F).

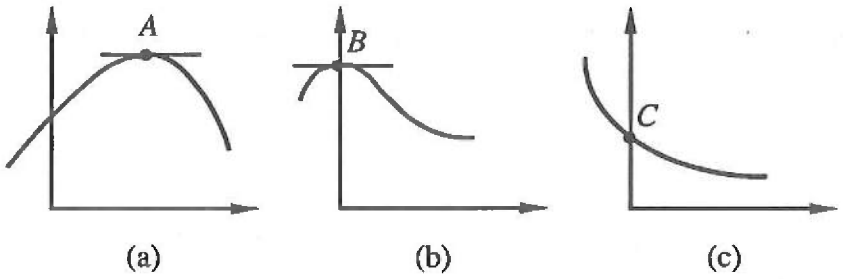


Figure 2.5 Conditions for a constrained maximum. Point A is an interior maximum; point B is a boundary maximum; point C is the constrained maximum by virtue of having the largest value in the feasible region.

The first type of solution, represented in Figure 2.5(a) as point A, has the maximum of F lying within the feasible region ($x_1 \geq 0$). In this case, $x_1 \neq 0$, so by condition (2.52) we require the first derivative to vanish ($u_1 = 0$). In the second type, known as a boundary solution, shown as point B in Fig. 2.5(b), F attains its maximum on the boundary. In this case both u_1 and x_1 are equal to zero and condition (2.52) is satisified. In the final possiblity, the solution may take the form of point C in Fig. 2.5(c), because to qualify as a solution the point need only have a value larger than any other point within the feasible region. This type of solution is characterized by a negative first derivative ($u_1 < 0$) and $x_1 = 0$. Again, the product condition (2.52) is satisfied.

Because the Kuhn-Tucker conditions are inequalities, they cannot be solved directly. One approach to finding the optimal solution is to note, according to conditions (2.51) - (2.53), that for each variable x_i, either $x_i = 0$ or $u_i = 0$ (or both). A solution can be sought by exploring the 2^n possible situations that result. For large n, of course, this approach is completely impractical. Several, more sophisticated, methods have therefore been developed.[5] The method employed in the experimental

portion of this chapter (Sec. 2.6) is known as the simplex method for quadratic programming. A brief description of the idea behind it follows; the complete development of the technique can be found in Ref. 6.

The Lagrangean $\phi(\mathbf{v},\mathbf{x})$ is simply the sum of the function F and a constraint term $-\mathbf{v}'(\mathbf{Ax}-\mathbf{b})$. If the constraint $\mathbf{Ax} = \mathbf{b}$ is satisfied then the Lagrangean is simply equal to F. We can therefore maximize $F(\mathbf{x})$ by increasing the value of $\phi(\mathbf{v},\mathbf{x})$ while maintaining the condition $\mathbf{Ax} = \mathbf{b}$. If we find that, for some $(\mathbf{v},\mathbf{x})$-pair, one of the u-variables is positive (i.e., the first derivative of ϕ is positive in some dimension), then an increase in the corresponding x-variable will cause an increase in ϕ. Upon raising the x-variable, the linear constraint will, in general, no longer be satisfied. It is necessary, therefore, to compensate by adjusting other x-variables to continue to satisfy the constraint. If this is achieved by changing x-variables that have no effect on ϕ (i.e., those for which the first derivative is zero), and if the change is made without changing the first derivatives, then each step of this type leads to a monotonic increase in ϕ while satisfying the constraint. Equivalently, each succeeding iteration of this kind produces an increase in F without violating the constraints.

2.5 Dimensionality effects and multiple-class sorting

As we have seen, the problem of classifying images in their vector representation reduces to one of suitably partitioning a space of very large dimension. While, in general, it is difficult to predict whether this or that algorithm will be successful in classifying a particular set of images, a few things can be said about the vector space that suggest rough conditions for linear separability and guidelines for multiple-class sorting.

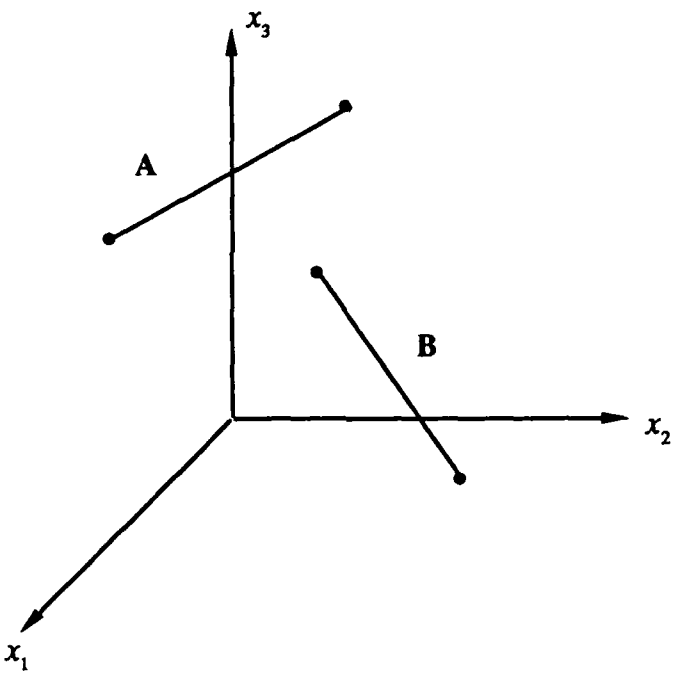


Figure 2.6 Convex hulls of classes containing two three-pixel images. In the ideal case, their intersection is a zero-probability event.

To begin, consider the hypothetical classification problem illustrated in Fig. 2.6. The two classes, labelled as "A" and "B", consist of two images composed of

three pixels each. The line segments in Fig. 2.6 represent the convex hulls of the two classes of image vectors. As we have shown in Section 2.2, the classes are separable if the convex hulls do not overlap. In this simple example, we see that, barring systematic relationships among the images, there is zero probability of overlap of the convex hulls for "A" and "B" (the line segments occupy no volume). Ideally, therefore, classes containing two arbitrary images in s-space are always separable. We can easily extend this result to N dimensions: two classes of N-pixel images are, in general, separable if they contain less than N images.

Unfortunately, we cannot directly apply this finding to real-world problems for several reasons. First, images are usually digitized at some stage in the process and are not free to take on the continuum of values suggested by Fig. 2.6. Second, real detection systems are subject to noise, which causes the convex hulls to have non-zero measure. Finally, images are not arbitrary points in space, but rather have systematic relationships among them. In particular, images within a class may be very much the same except for minor features that distinguish them. When that is the case, they may lie in a small subspace of the space defined by the pixels.

The net effect of these considerations is to weaken our previous assertion concerning the effect of dimensionality on class separability. Without reference to a particular set of images, what can safely be said is that two image classes having significant variation in M pixels are likely to be linearly separable, provided that there are less than M images in each class.

This suggests large classes of unrelated images can be distinguished. It may, therefore, be possible to merge classes into superclasses to facilitate multiple-class

sorting. With proper grouping of the classes, discriminant vectors can be designed to bisect the set of classes with each decision. In this way, an image can be associated with one of K classes in $c(\log_2 K)$ binary decisions, where $c(x)$ denotes the smallest integer greater than or equal to x. This approach demands, of course, that the superclasses, as well as the individual classes, be linearly separable.

An alternative is pairwise, sequential elimination of the classes. This requires (K-1) two-class decisions to solve a K-class problem. Classification in a three-class problem (classes A, B, and C), for example, might go as follows. Apply the A-versus-B discriminant vector to the input image. If the decision is for B, apply the B-versus-C discriminant vector. If the decision is for C, then we have classified the image as an element of class C in two [i.e., (K-1)] steps.

2.6 Experimental results

Two sets of experiments were performed to demonstrate the convex-hull separation approach. First, a two-character recognition problem was constructed to test the method in its basic form. Second, an expanded problem was considered in which eight characters were incorporated.

Examples of the images used in the two-class problem are shown in Fig. 2.7. In the actual experiments, seven fonts were used for training and five were reserved as test examples. The images contain 64 × 64 (= 4K) pixels and although they are essentially binary, 256 grey levels were possible. Each of the images was normalized so that its vector was of unit length.

The normalized training images took the place of the **x**-vectors in Eq. (2.30) and on this basis, the quadratic programming problem of Eqs. (2.41)-(2.43) was constructed. A FORTRAN implementation of the simplex method for quadratic programming, included in Appendix B, was written and was used to solve for the λ-coefficients that minimize the distance function d^2. The optimum λ-coefficients were then substituted into Eq. (2.30) to find the points of closest approach of the convex hulls of the two classes. The normalized difference vector of these two points acts as the discriminant vector in the classification problem (see Fig. 2.2). The discriminant vector components were rearranged into the two-dimensional image format; the resulting array appears in Fig. 2.8. The dark portions of the array represent negative values of the discriminant vector; the light areas signify positive values.

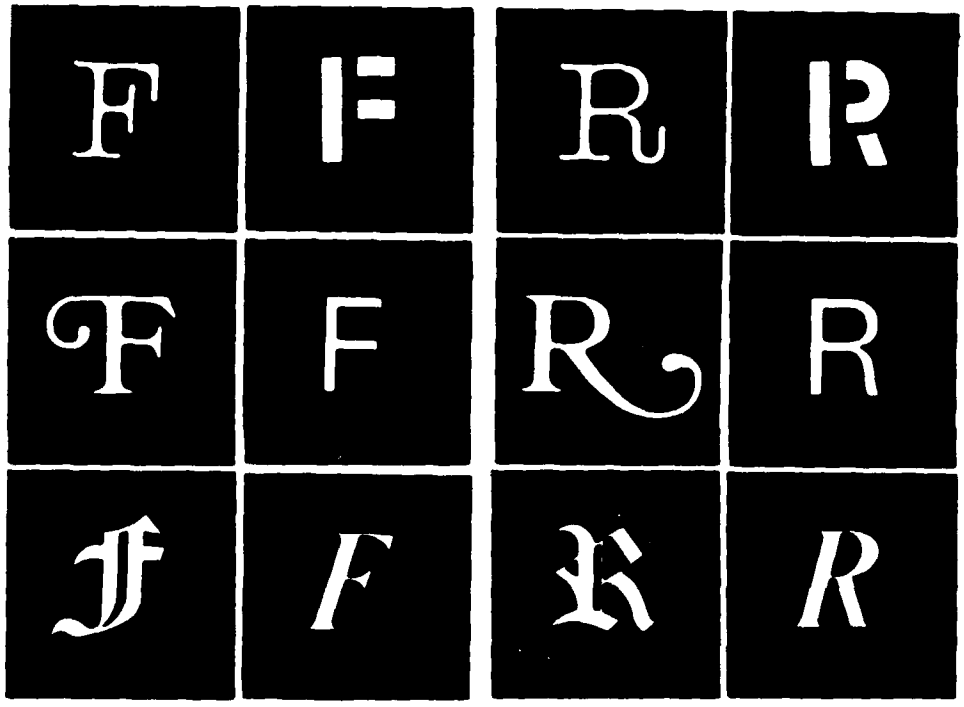


Figure 2.7 Examples of character images used in the first experiment. Seven fonts were used as prototypes; five fonts were reserved as test examples.

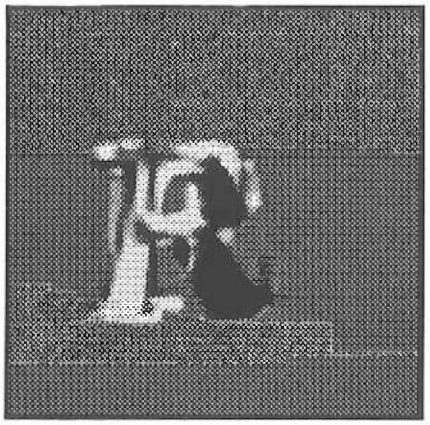


Figure 2.8. Two-dimensional representation of convex-hull discriminant vector for "F" and "R."

To provide a basis for comparison for the convex-hull approach, discriminant vectors were computed in accordance with two other methods: the Fukunaga-Koontz transform and the difference of means.

The generation of the Fukunaga-Koontz basis vectors followed the technique described in Ref. 8. Based on Eq. (1.14), the quantity $(\lambda_i - 0.5)$ can be defined as a separation power for Fukunaga-Koontz basis vectors.[8] In theory, those basis vectors for which the separation power is largest are most useful for discrimination of the classes. By this prescription, the two best basis vectors were chosen. A linear combination of these vectors was ordered into a two-dimensional image format; the result appears in Fig. 2.9. The difference-of-means discriminant vector is similarly displayed in Fig. 2.10.

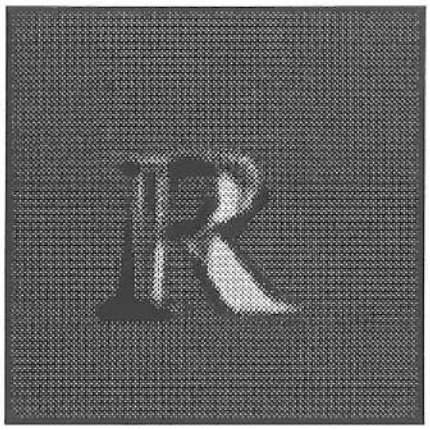


Figure 2.9 Two-dimensional representation of Fukunaga-Koontz discriminant vector for "F" and "R."

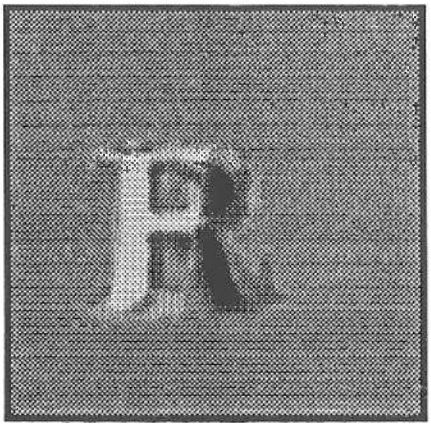


Figure 2.10 Two-dimensional representation of difference-of-means discriminant vector for "F" and "R."

The results of the two-character classification experiment were computer-simulated by forming the inner product between the character images and the discriminant vectors. Two criteria were used to compare the convex-hull separation approach to the other two approaches. The first criterion is the minimum separation of

the projections of the images onto the discriminant vector (remember that the convex-hull approach maximizes this criterion). The second criterion, known as the d'-parameter, defined as

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1 \sigma_2}} \quad , \tag{2.54}$$

is a traditional statistical pattern recognition measure of average separation of the classes.[9] In Eq. (2.54), μ_j is the mean of the projections of the images from class j onto the discriminant vector and σ_j is the standard deviation. The results for the minimum separation and d'-criterion appear in Tables 2.1 and 2.2.

Note that in addition to the convex-hull, difference-of-means, and Fukunaga-Koontz approaches, two other discriminant vectors are included in the comparisons: "F minus R" (the difference of a single F and a single R from the training set), and "F" (a particular F from the training set). Of course, these methods are not recommended, but it is instructive to consider their performance.

Table 2.1 Minimum Separation of Projections for Various Discriminant Vectors

Discriminant vector	Training images	Test images	Overall
Convex hull	0.391	0.284	0.284
Difference of means	0.131	0.308	0.131
Fukunaga-Koontz	0.062	0.071	0.062
F minus R	0.013	0.119	0.013
F	- 0.216	-0.098	-0.301

Table 2.2 Values of the d'-parameter for Various Discriminant Vectors

Discriminant vector	Training images	Test images	Overall
Convex hull	793.12	8.32	11.74
Difference of means	4.69	5.77	4.68
Fukunaga-Koontz	3.64	3.82	3.77
F minus R	2.91	5.00	3.16
F	1.49	1.87	0.28

Note that, except for the single-F discriminant vector, each method successfully classifies all of the images tested, however the convex-hull method provides the best performance by both criteria. For a point of reference, it should be noted that normalized images lie exclusively in the non-negative orthant of the unit hypersphere, therefore the maximum value that the minimum separation criterion can attain is $\sqrt{2}$.

In the second experiment, an expanded character set was used. Here, eight characters (A, B, C, D, U, O, F, and R) constituted eight image classes. Examples of these images are shown in Fig. 2.11. Ten fonts were assembled as examples of each character so that a total of 80 images were included in the test. As discussed in the previous section, two paths to multiple-class sorting can be taken. We can achieve the desired result either by pursuing a strategy of pairwise elimination of the classes (K-1 decisions for K classes) or by repeatedly bisecting the decision space [int($\log_2 K$)+1 decisions for K classes]. We have already considered the two-class decision that forms the building block of the pairwise elimination approach. We now examine the latter, more efficient approach.

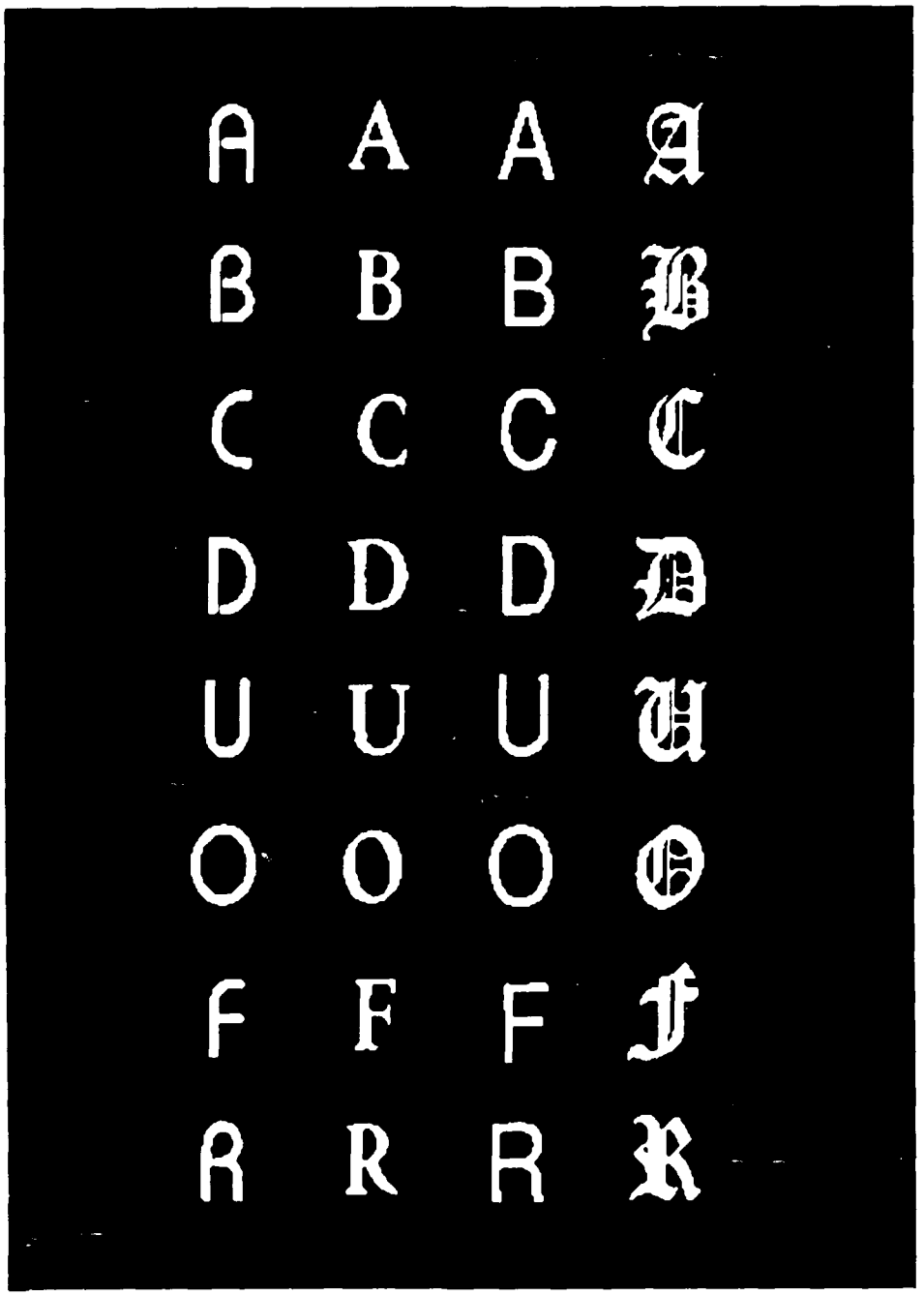


Figure 2.11 Examples of images in eight-class problem. Classes consisted letters A, B, C, D, U, O, F, and R in ten different fonts.

In this experiment, $\log_2 8$ (=3) binary decisions are used to classify eight characters:

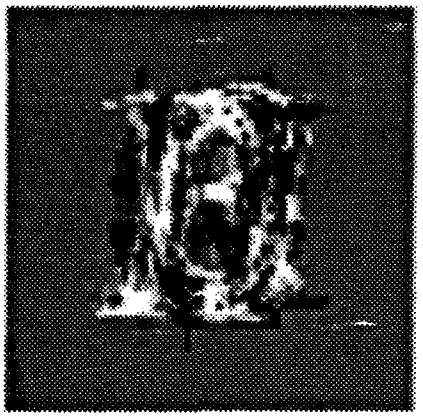
1. ABCD - UOFR
2. ABUO - CDFR
3. ADUR - BCOF

In this decision structure, if the two possible outcomes of each decision are labeled as 0 and 1, we see that each class is represented by a unique binary word in terms of the three decision outcomes:

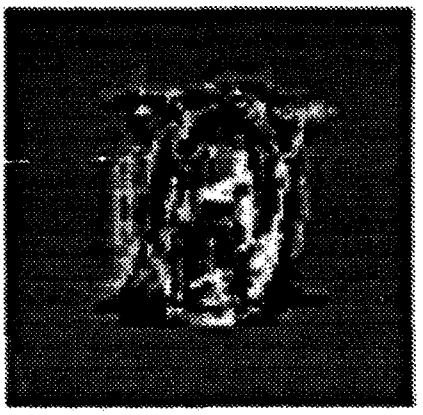
A	0	0	0
B	0	0	1
C	0	1	1
D	0	1	0
U	1	0	0
O	1	0	1
F	1	1	1
R	1	1	0

The discriminant vectors for these decisions, computed as in the previous experiment, by the simplex method, are shown in Fig. 2.12. Again, the results obtained using these discriminant vectors were simulated. Figure 2.13 is a scatter plot of the inner product results. Each axis represents the inner product with one of the three discriminant vectors. In the simulation, it was found that all 80 characters were correctly classified.

(a)



(b)



(c)

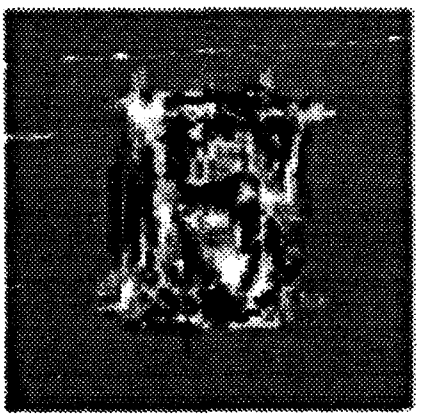


Figure 2.12 Two-dimensional representations of convex-hull discriminant vectors for eight-class problem: (a) ABCD-UOFR; (b) ABUO-CDFR; (c) ADUR-BCOF.

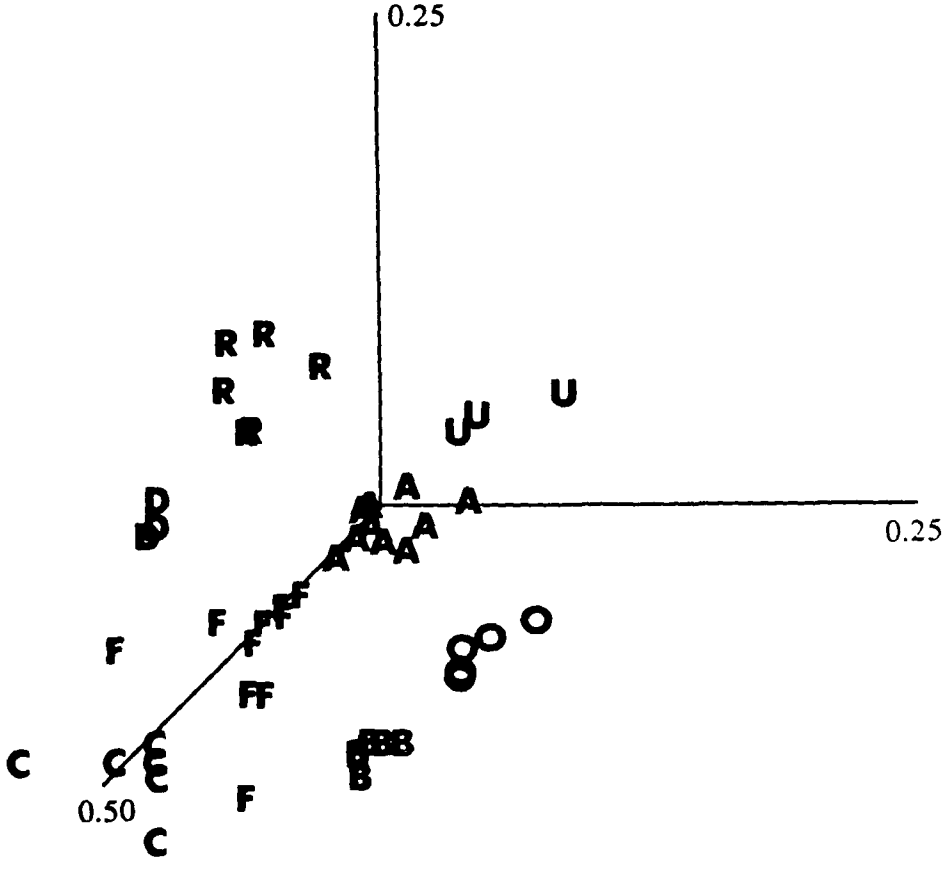


Figure 2.13 Scatter plot of inner products of 80 characters with three convex-hull discriminant vectors.

References for Chapter 2

1. J. Van Tiel, *Convex Analysis: An Introductory Text*, (Wiley, New York, 1984).
2. V. G. Boltyanskii, *Optimal Control of Discrete Systems*, (Wiley, New York, 1984).
3. J. B. Rosen, "Pattern Separation by Convex Programming," J. Math. Anal. and Appl **10**, 123-124 (1965).
4. K. C. Border, *Fixed point theorems with applications to economics and game theory*, (Cambridge University Press, Cambridge, 1985).
5. G. B. Dantzig, *Linear Programming and Extensions*, (Princeton University Press, Princeton, 1963).
6. C. Van de Panne and A. Whinston, "The Simplex and the Dual Method for Quadratic Programming," Operat. Res. Quart., **15**, 4, 355-388 (1964).
7. A. C. Chiang, *Fundamental Methods of Mathematical Economics*, (McGraw-Hill, New York, 1984).
8. J. R. Leger and S. H. Lee, "Image classification by an optical implementation of the Fukunaga-Koontz transform," J. Opt. Soc. Am. **72**, 556-564 (1982).
9. H. H. Barrett and W. Swindell, *Radiological Imaging*: *The Theory of Image Formation, Detetection, and Processing*, v. 2, (Academic Press, New York, 1981).

Chapter 3:
Classification of quantum-limited images

In this chapter, we examine the classification of quantum-limited, as opposed to classical intensity, images. The purpose of considering quantum-limited images is twofold. First, it may provide useful insights for naturally quantum-limited applications such as low-dose radiological imaging and electron microscopy, astronomy, and night vision. Second, certain properties of the acquisition process for quantum-limited images suggest that these images might be utilized to advantage, even when there is an abundance of available light.

A quantum-limited image, acquired using a position-sensitive photon-counting system, can be represented by a list of the digitized spatial coordinates of the photoevents of which it is composed. In contrast, a digitized, classical intensity image is specified by a list of (typically, on the order of a million) pixel intensity values. The quantity of data needed to specify a classical intensity image requires that a costly dimensionality reduction step be performed before a recognition or classification decision can be made by digital means. The promise for quantum-limited image classification lies in eliminating this dimensionality reduction step in the course of acquiring the image. If a quantum-limited version of some image, described by, say, a few thousand photoevent locations, contains sufficient information to classify it, then a tremendous computational savings has been made by

way of the detection process. Experiments described in subsequent sections demonstrate that quantum-limited images do, in fact, provide this sort of data compression advantage.

In addition to the interesting computational properties afforded by the use of quantum-limited images, the acquisition system from which they are derived has several appealing features. The system, when packaged as a camera, can be quite compact and sturdy. Unlike traditional optical pattern recognition systems, photon-limited imaging systems operate in white light, avoiding the complication of the incoherent-to-coherent conversion step inherent in the standard optical correlator approach. Further, though the detection process is optical, the computations are performed digitally, thus avoiding the flexibility problems associated with storing prototype images and/or discriminant vectors in holographic frequency plane filters. Finally, the photon-counting imaging system operates in what one would normally consider total darkness — an obvious advantage in many applications.

The focus of this chapter is to demonstrate the potential for utilizing photon-counting systems to classify images. The chapter begins, in Section 3.1, with an overview of photon-counting systems and a brief description of the particular system used in the present research. Section 3.2 describes the statistical properties of the images produced by these systems. In Section 3.3, photon correlation methods are reviewed and the quantum-limited inner product is introduced. The statistical

several image classification solutions are derived in Section 3.6. In Section 3.6.1, these solutions and standard pattern recognition methods are implemented experimentally and the results compared. A method for classifying images subject to in-plane rotations is proposed in Section 3.7 and experimentally demonstrated in Section 3.7.1. The chapter concludes with a discussion, in Section 3.8, of how the statistical decision theory solutions described in Section 3.6 might be applied to the classical-intensity image classification problem. Simulations of the results of such an experiment are presented in Section 3.7.1.

3.1 Photon-counting imaging systems

Most two-dimensional photon-counting systems described in the literature use a photocathode, a set of microchannel plates,[1,2] and an anode assembly to determine the locations of detected photoevents. The primary difference between various systems lies in the design of the anode. Anode structures that have been reported include silicon-intensified-target television cameras,[3-5] self-scanned CCD arrays,[6-11] crossed-wire-grid anodes,[12] multi-anode arrays,[13-15] wedge-and-strip anodes,[16-21] grey-coded masks used with a bank of photomultipliers,[22,23] and resistive anodes.[24-30]

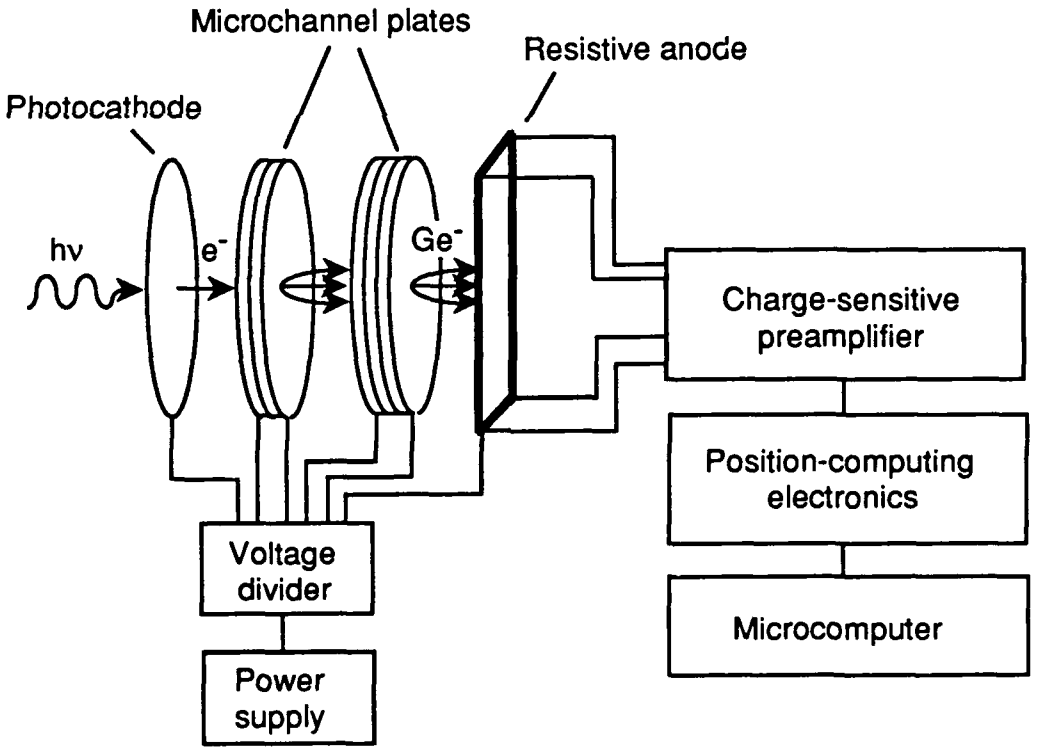


Figure 3.1 Schematic diagram of photon-counting system.

The device employed in the present research uses a resistive anode to provide photoevent position information.[31] A schematic diagram of the system is illustrated in Fig. 3.1. A photon, incident on the photocathode, causes the emission of a photoelectron. The electron is accelerated by an applied field across a small gap toward the first in a stack of microchannel plates. The plates are thin glass discs perforated by small parallel pores or channels, coated with a semiconductor material. The channels act as tiny photomultipliers, each producing a cascade of electrons. To increase gain, the channels lie at an angle to the direction of the applied field. To prevent ion feedback, the plates are assembled to form V- and Z-configurations of the channels. The complete plate assembly produces an electron gain of approximately 10^7. The charge packet that emerges at the output of the stack is deposited on the anode, consisting of a sheet of resistive material, bounded by four circular arcs, to which an electrode is attached at each of its four corners. The amount of charge reaching the electrodes along one edge of the anode is directly proportional to the distance from the point of impact of the charge pulse to the opposite edge. The desired photoevent location can, therefore, be deduced by computing a centroid of the charges measured at the electrodes. The spatial coordinates of each detected photoevent are computed by dedicated analog hardware,[32] then digitized and passed to a microcomputer for processing and/or display.

To permit its convenient use as an imaging device, and to protect it from excessive light levels, the detector used in this research was packaged in an aluminum housing as a camera. The housing, shown in Fig. 3.2, was designed by the author and his colleagues, Anthony Martino and Thomas Isberg. At the front of the housing is a

standard bayonet mount for the lenses used in 35mm-format photography. Between the lens and the photocathode is a carriage that can be moved into one of two positions. In one position, a mirror contained in the carriage diverts the image to a focusing screen where it can be examined with the attached eyepiece. In the other position, a stack of neutral density filters intervenes between the lens and photocathode, thus producing an acceptable light level for the detector.

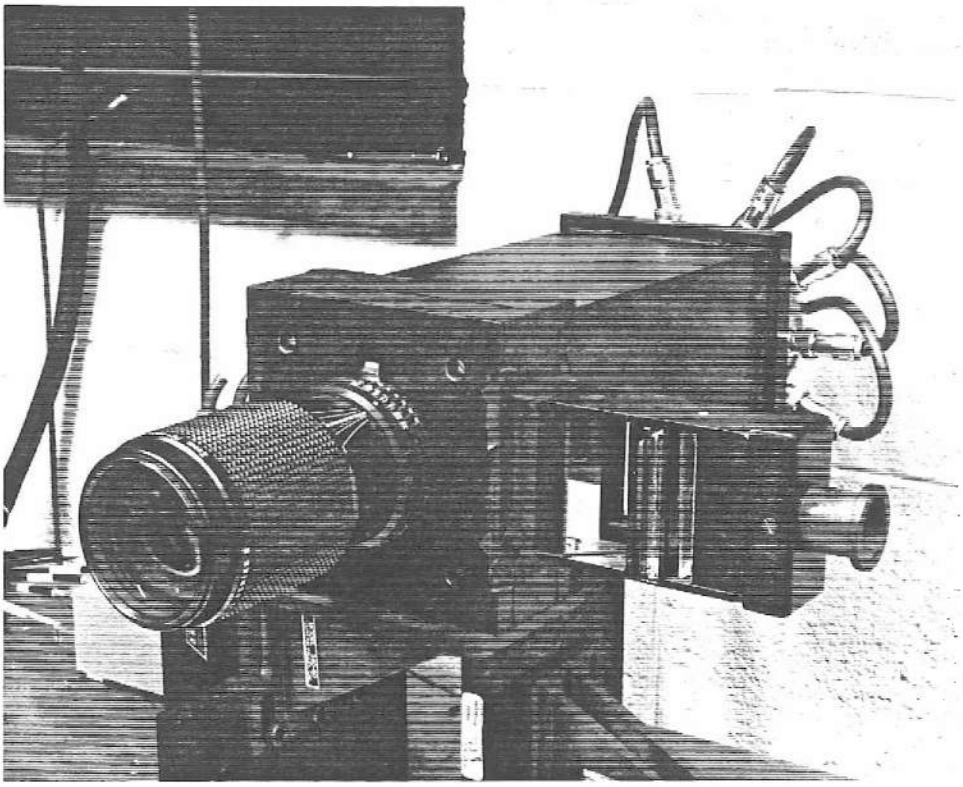


Figure 3.2 Housing for photon-counting detector.

The present system reports one photoevent at a time and can operate at a rate of about 10^5 counts/sec. For high speed applications, anodes that detect many events

simultaneously might be preferable. Such systems are unsuitable for some applications, however, since they fail to provide time-of-arrival information and do not permit precise control of the number of photoevents. Instead, such a system is run for a predetermined integration time, the number of detected photoevents becomes a random variable, and the arrival times of particular events cannot be deduced.

In the next section, the statistical properties of the images produced by photon-counting systems are reviewed in terms of both the fixed-integration-time and fixed-photoevent-count photon-counting configurations.

3.2 Statistics of the quantum-limited image

A digitized, classical-intensity image is quantized in two ways. First, it is integral sampled, dividing the image plane into picture elements (pixels). Second, the averaged intensities are quantized for digital processing. In contrast, a digitized, photon-limited image is artificially quantized only in the former sense. The value assigned to each pixel is the number of photoevents detected within that pixel. Hereinafter, a photon-limited image is denoted by a vector $\mathbf{n}$, the components of which are the pixel photoevent counts. The image is assumed digitized to produce N possible photoevent locations (pixels) and the count for pixel i is denoted by n_i, the ith component of $\mathbf{n}$.

a) Fixed-integration-time configuration

If a photon-limited image is acquired over a fixed integration time, τ, then both the pixel photoevent counts, n_i ($i = 1, 2, \ldots, N$), and the total number of detected photoevents, N_t, are random variables. Treating the pixels of the photon-limited imaging system as detectors occupying non-overlapping areas A_i, and supposing an image irradiance $I(x,y;t)$, the photoevent count for pixel i during a time interval $[t, t+\tau]$ is described by an inhomogeneous Poisson process,[33-37]

$$P[n_i|I(x,y;t)] = \frac{\left[-\int_t^{t+\tau} dt' \int_{A_i} dxdy\lambda(x,y;t')\right]^{n_i}}{n_i!} \cdot \exp\left[-\int_t^{t+\tau} dt' \int_{A_i} dxdy\lambda(x,y;t')\right] . \tag{3.1}$$

In Eq. (3.1) the rate function $\lambda(x,y;t)$ is given by

$$\lambda(x,y;t) = \frac{\eta I(x,y;t)}{h\nu} \quad , \tag{3.2}$$

where η is the quantum efficiency of the detector, h is Planck's constant, and ν is the mean frequency of the incident quasi-monochromatic light.

In general, $I(x,y;t)$ is a random process and the observable counting distributions, $p(n_i)$, are obtained by taking an average of Eq. (3.1) over an ensemble of integration times. In the case in which the image is composed of polarized, quasi-monochromatic thermal light, if the integration time, τ, is small compared with the coherence time of the source, then it can be shown that the n_i obey the Bose-Einstein distribution.[37] Of more practical significance for the present research is the case in which the integration time is, instead, much larger than the coherence time. In this case, the irradiance can be regarded as constant in time [$I(x,y;t) = I(x,y)$]and the photoevent count for pixel i obeys the Poisson distribution,[35]

$$p(n_i) = \frac{< n_i >^{n_i}}{n_i!} \exp(- < n_i >) \quad , \tag{3.3}$$

where

$$< n_i > = \frac{\eta\tau}{h\nu} \int_{A_i} dxdy\, I(x,y) \tag{3.4}$$

is the mean photoevent count for pixel i. Incidentally, the same result is obtained when the irradiance does not fluctuate significantly, as in the case of illumination provided by a well-stabilized single-mode laser.[37]

Neglecting dead-time effects in the detector, the pixel photoevent counts can be taken to be independent random variables, and the joint density for the n_i, denoted by $p(\mathbf{n})$, is simply the product of the individual distributions:

$$p(\mathbf{n}) = \prod_{i=1}^{N} \frac{< n_i >^{n_i}}{n_i!} \exp(- < n_i >) \quad . \tag{3.5}$$

In this case, since the total photoevent count is the sum of the pixel photoevent counts, the density function for the total count is simply the convolution of the pixel count distributions, $p(n_i)$. The result is that the total photoevent count, N_t, also obeys the Poisson distribution,

$$p(N_t) = \frac{< N_t >^{N_t}}{N_t!} \exp(- < N_t >) \quad , \tag{3.6}$$

where

$$< N_t > = \sum_{i=1}^{N} < n_i > \quad . \tag{3.7}$$

Using Eqs. (3.4) and (3.7), an alternative expression for the mean pixel photoevent counts is obtained:

$$< n_i > = < N_t > \frac{\int_{A_i} dxdy\, I(x,y)}{\sum_{i=1}^{N} \int_{A_i} dxdy\, I(x,y)} \quad . \tag{3.8}$$

The integrated intensities in Eq. (3.8) are simply the classical pixel intensity values, referred to in Chapter 1 as x_i, that result from integral sampling of the image. Equating the integrated intensities with the x_i, we obtain the relationship between the mean pixel photoevent counts and the classical pixel intensities, x_i,

$$< n_i > = < N_t > \frac{x_i}{\sum_{i=1}^{N} x_i} \quad . \tag{3.9}$$

b) Fixed-photoevent-count configuration

An alternative to counting photons for a fixed length of time is to count up to some fixed number, N_t. This process is characterized by N_t trials (the photoevent detections) having one of N possible mutually exclusive outcomes (the N possible photoevent locations). In the absence of dead-time effects, successive photoevent locations are independent and the pixel photoevent counts obey the multinomial distribution

$$p(\mathbf{n}) = N_t ! \prod_{i=1}^{N} \frac{p_i^{n_i}}{n_i !} \quad , \tag{3.10}$$

where p_i is the probability that a particular photoevent is detected by pixel i, given by[37]

$$p_i = \frac{\int_{A_i} dxdy\, I(x,y)}{\sum_{i=1}^{N} \int_{A_i} dxdy\, I(x,y)} = \frac{x_i}{\sum_{i=1}^{N} x_i} \quad . \tag{3.11}$$

The mean photoevent count for pixel i is given by

$$< n_i >= N_t \frac{x_i}{\sum_{i=1}^{N} x_i} \quad . \tag{3.12}$$

3.3 Photon correlation and the quantum-limited inner product

Recently, photon correlation techniques have received considerable attention. The triple correlation has been shown to permit photon-limited imaging in the presence of atmospheric turbulence[38-40] and random shifts and rotations of the image.[41-43] Double correlation techniques (auto- and cross-correlation of photon-limited images) provide information about object dimensions, velocity, and other parameters.[44] Until recently, the correlation between a quantum-limited image and a fixed reference function was not considered. The value at the origin of such a correlation function is the inner product that forms the backbone of the pattern recognition approaches described in Chapter 1.

Rose[45] first discussed the identification of quantum-limited images in terms of human observer performance. He demonstrated that a relatively small number of photoevents is sufficient for a human to observe intensity variations in an image. Saxton and Frank[46] have suggested that image correlation can be used to identify motifs in low-dose electron microscopy. Morris developed the statistical properties and implementation procedure for quantum-limited matched filtering[47] and demonstrated in computer simulations[48] that a small number of photons can be sufficient for machine-based recognition of images. Morris, Wernick, and Isberg[49] performed laboratory experiments that verified the results of the computer simulations.

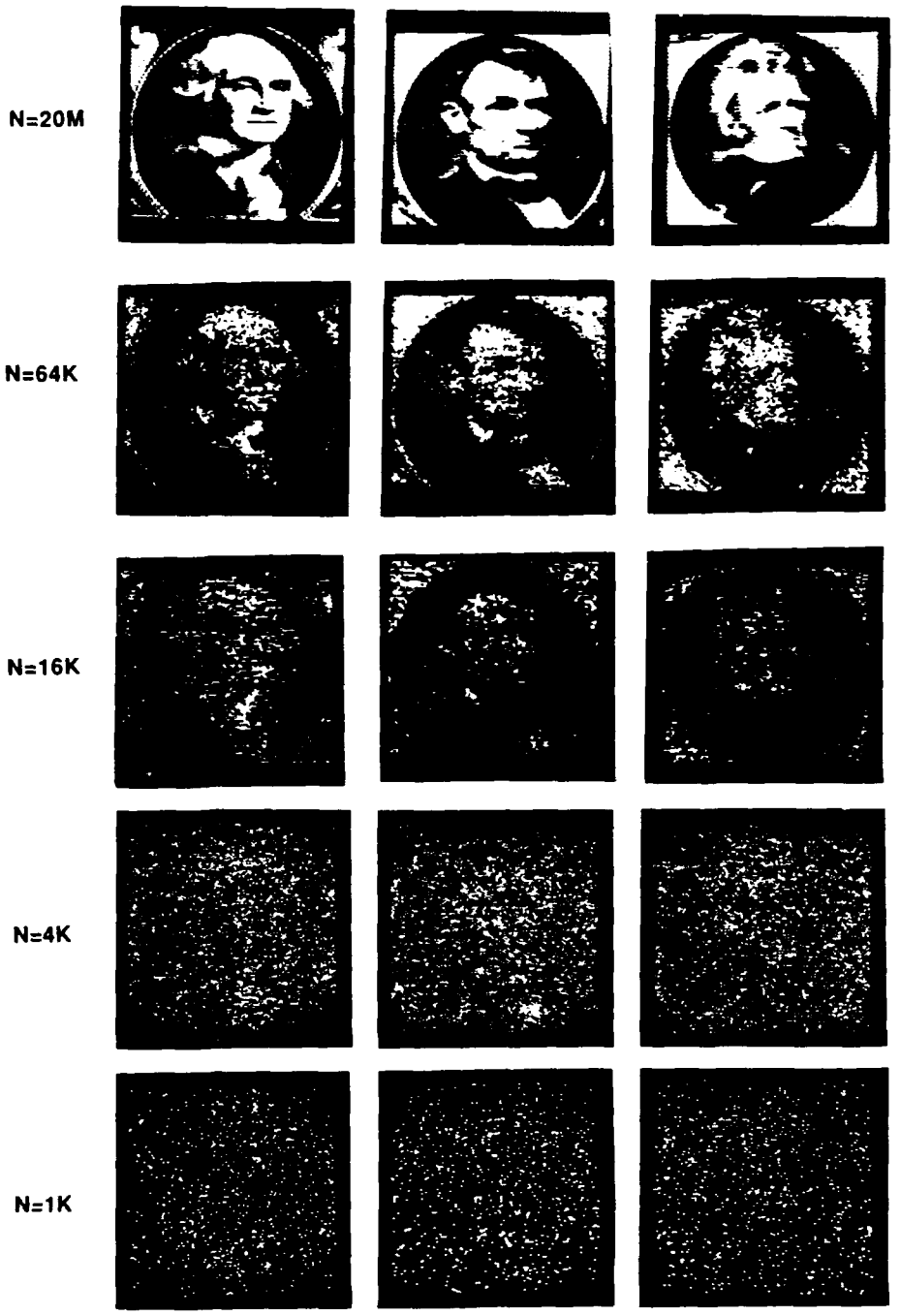


Fig. 3.3 Images of portraits made with photon-counting detection system. Number of photoevents constituting images in each row is indicated at left.

Examples of the sort of images considered in these experiments are illustrated in Fig. 3.3. The images are arranged, from top to bottom, in order of decreasing number of photoevents. Each of these images can be described by a list of digitized spatial (x,y) coordinates of the detected photoevents. Equivalently, these photoevent locations can be used to form the pixel-photoevent-count vector, **n**, introduced in the previous section.

In Chapter 1, the inner product, C, between an input image vector **x** and a discriminant vector **h**, given by

$$C = \sum_{i=1}^{N} x_i h_i \quad , \tag{3.13}$$

was introduced as the central quantity in a number of image recognition and classification schemes. In Eq. (3.13), i is the pixel index and N is the total number of pixels in the image. In this chapter, we consider the inner product between a photon-limited version of the same image (described by the histogram vector **n**) and the discriminant vector **h**. Hereinafter, this quantity will be termed the quantum-limited inner product and denoted by C_{QL}. Because the quantum-limited image is the result of a stochastic process, C_{QL}, given by,

$$C_{QL} = \sum_{i=1}^{N} n_i h_i \quad , \tag{3.14}$$

is a random variable.

One of the important features of this quantity, demonstrated in the following section is that its expected value is directly proportional to the inner product between **h** and **x** (the classical intensity image underlying **n**). In effect, therefore, the inner

product obtained through use of the quantum-limited version of an image provides a Monte Carlo estimate of the classical-intensity inner product given in Eq. (3.13). Hence, the quantum-limited image can be directly applied, in lieu of the classical-intensity image, in any of the pattern recognition approaches described in Chapter 1.

The effectiveness of utilizing the quantum-limited inner product when the classical-intensity image is available can be seen by considering the manner in which C_{QL} is calculated. As discussed in Section 3.1, many photon-counting imaging systems report the location of detected photoevents, one event at a time. For this type of system, an alternative, computationally efficient expression for C_{QL} can be obtained. The pixel-photoevent-count for pixel i can be represented as a sum of Kronecker delta functions,

$$n_i = \sum_{p=1}^{N_t} \delta_{i i_p} \quad , \tag{3.15}$$

where i_p denotes the measured pixel location of the pth photoevent. Substituting from Eq. (3.15) into the definition of C_{QL} [Eq. (3.14)], we obtain

$$C_{QL} = \sum_{i=1}^{N} \sum_{p=1}^{N_t} \delta_{i i_p} h_i \quad , \tag{3.16}$$

or, using the sifting property of the delta function,

$$C_{QL} = \sum_{p=1}^{N_t} h_{i_p} \quad . \tag{3.17}$$

The procedure for calculating C_{QL} indicated by Eq. (3.17) is as follows. The discriminant vector **h** is loaded into computer memory to prepare the system. An image to be recognized or classified is then presented to the detector. During the

operation of the detector, a stream of photoevent locations is reported to the computer. As each photoevent location is received, the value, at that pixel, of the discriminant vector **h** is retrieved from memory and added to an accumulated sum. The procedure is repeated for each detected photon until the experiment is concluded (signalled, depending on the experimental configuration, by the end of the preselected integration time or by the counting of the specified number of photoevents). The value of the accumulated sum at the conclusion of the experiment is equal to C_{QL}.

The computation of C_{QL} according to this method requires N_t additions which is usually much less than the N multiplications and N additions required by Eq. (3.14). Further, the additions entailed in the proposed method are performed concurrently with the acquisition of the image, hence the image has been analyzed by the time it has been acquired. The current generation of photon-limited image acquisition systems operate at a rate on the order of 10^6 counts/sec. If a reliable inner product estimate can be achieved by computing C_{QL} with, say, a few thousand photoevents (as is the case for the images shown in Fig. 3), then an image can be acquired and identified in a matter of milliseconds.

To evaluate the reliability of the estimate provided by the quantum-limited inner product, C_{QL}, it is necessary understand its statistical properties. These properties have important significance for system design, in particular for construction of decision boundaries and for calculation of the number of photoevents required to achieve the desired reliability level. In the next section, the relevant statistics of C_{QL} are derived and the proportionality between the average of C_{QL} and its classical-intensity counterpart, C, is demonstrated.

3.4 Statistics of the quantum-limited inner product

It can be shown[47] that the quantum-limited inner product, C_{QL}, obeys the Gaussian distribution

$$p(C_{QL}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(C_{QL} - < C_{QL} >\right)^2}{2\sigma^2} \right] \quad . \tag{3.18}$$

The mean value of the quantum-limited inner product, C_{QL}, is given by

$$< C_{QL} >= \sum_{n_1=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} p(\mathbf{n}) C(\mathbf{n}) \quad , \tag{3.19}$$

where the <...> represents an average over an ensemble of realizations of the photon-limited image, **n**. Equation (3.19) can be written alternatively as

$$< C_{QL} >= \sum_{i=1}^{N} h_i \sum_{n_i=0}^{\infty} n_i p(n_i) \quad , \tag{3.20}$$

or equivalently,

$$< C_{QL} >= \sum_{i=1}^{N} h_i < n_i > \quad . \tag{3.21}$$

Substituting from Eq. (3.9) for $< n_i >$, the mean value of C_{QL} for the fixed-integration-time experiment becomes

$$< C_{QL} >= \left(\frac{< N_t >}{\sum_{i=1}^{N} x_i} \right) \sum_{i=1}^{N} x_i h_i \quad . \tag{3.22}$$

Identifying the inner product term as the classical-intensity inner product, C, we obtain the proportionality relationship between C and the ensemble average of C_{QL},

$$< C_{QL} >= \left(\frac{< N_t >}{\sum_{i=1}^{N} x_i} \right) C \quad . \tag{3.23}$$

For the fixed-photoevent-count experiment, a similar result is obtained:

$$< C_{QL} >= \left(\frac{N_t}{\sum_{i=1}^{N} x_i} \right) C \quad . \tag{3.24}$$

Note that the mean of the photoevent count, $< N_t >$, that appears in the fixed-time expression [Eq.(3.23)] is replaced by N_t in the fixed-count result [Eq. (3.24)].

The variance of the inner product, defined in the usual way as

$$\sigma^2 = < C_{QL}{}^2 > - < C_{QL} >^2 \quad , \tag{3.25}$$

can be derived in a similar manner. The term, $< C_{QL}{}^2 >$, is given by

$$< C_{QL}{}^2 >= \left\langle \left(\sum_{i=1}^{N} n_i h_i \right)^2 \right\rangle \quad , \tag{3.26}$$

or,

$$< C_{QL}{}^2 >= \left\langle \sum_{i=1}^{N} \sum_{i'=1}^{N} n_i n_{i'} h_i h_{i'} \right\rangle \quad . \tag{3.27}$$

Bringing the ensemble averaging operation inside the summation, and excluding from it the deterministic h_i terms, Eq. (3.27) becomes:

$$< C_{QL}{}^2 >= \sum_{i=1}^{N} \sum_{i'=1}^{N} < n_i n_{i'} > h_i h_{i'} \quad . \tag{3.28}$$

The $< C_{QL} >^2$ term in Eq. (3.25) is given by

$$< C_{QL} >^2 = \left\langle \sum_{i=1}^{N} n_i h_i \right\rangle^2 \quad , \tag{3.29}$$

or

$$< C_{QL} >^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} < n_i >< n_{i'} > h_i h_{i'} \quad . \tag{3.30}$$

Substituting from Eqs. (3.28) and (3.30) into the expression for the variance of C_{QL} [Eq. (3.25)] and collecting the summations, we obtain

$$\sigma^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} (< n_i n_{i'} > - < n_i >< n_{i'} >) h_i h_{i'} \quad . \tag{3.31}$$

The term in parentheses in Eq. (3.31) is simply the covariance of the photoevent counts detected by pixels i and i'. Denoting this pixel-photoevent-count covariance as $\sigma_{ii'}$, Eq. (3.31) becomes

$$\sigma^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} \sigma_{ii'} h_i h_{i'} \quad . \tag{3.32}$$

In the fixed-integration-time experiment described in previous sections, the pixel-photoevent-counts are independent and Poisson-distributed. The pixel-photoevent-count covariance is therefore given by

$$\sigma_{ii'}{}^2 = \begin{cases} < n_i > & ; i = i' \\ 0 & ; i \neq i' \end{cases} \quad . \tag{3.33}$$

The value of $\sigma_{ii'}$ for the case $i = i'$ is a well-known property of the Poisson-distribution. The zero value for the case $i \neq i'$ is a result of the independence of the pixels.

Substituting from Eq (3.33) into Eq. (3.32) we obtain the result for the variance of the quantum-limited inner product for the fixed-integration-time experiment,

$$\sigma^2 = \sum_{i=1}^{N} h_i^2 < n_i > \quad , \tag{3.34}$$

or, using the expression for $< n_i >$ [Eq. (3.9)],

$$\sigma^2 = < N_t > \frac{\sum_{i=1}^{N} x_i h_i^2}{\sum_{i=1}^{N} x_i} \quad . \tag{3.35}$$

In the fixed-photoevent-count experiment, the pixel-photoevent-counts obey the multinomial distribution, and the covariance, $\sigma_{ii'}$, is given by (see Appendix C)

$$\sigma_{ii'} = \begin{cases} N_t p_i (1 - p_i) & ; i = i' \\ -N_t p_i p_{i'} & ; i \neq i' \end{cases} \quad . \tag{3.36}$$

Substituting from Eq. (3.36) into Eq. (3.32) we obtain

$$\sigma^2 = N_t \sum_{i=1}^{N} p_i (1 - p_i) h_i^2 - N_t \sum_{\substack{i,i'=1 \\ i \neq i'}}^{N} p_i p_{i'} h_i h_{i'} \quad . \tag{3.37}$$

Substituting from Eq. (3.12) for p_i, Eq. (3.37) becomes

$$\sigma^2 = N_t \frac{\sum_{i=1}^{N} x_i h_i^2}{\sum_{i=1}^{N} x_i} - N_t \frac{\sum_{i=1}^{N} x_i^2 h_i^2}{\left(\sum_{i=1}^{N} x_i \right)^2} - N_t \frac{\sum_{\substack{i,i'=1 \\ i \neq i'}}^{N} x_i x_{i'} h_i h_{i'}}{\left(\sum_{i=1}^{N} x_i \right)^2} \quad , \tag{3.38}$$

or, collecting the last two terms,

$$\sigma^2 = N_t \frac{\sum_{i=1}^{N} x_i h_i^2}{\sum_{i=1}^{N} x_i} - N_t \left(\frac{\sum_{i=1}^{N} x_i h_i}{\sum_{i=1}^{N} x_i} \right)^2 \quad . \tag{3.39}$$

Equation (3.39) represents the variance of the quantum-limited inner product for the fixed-photoevent-count experiment. Note that the first term in the expression is identical to the variance in the fixed-time configuration if N_t is replaced by $< N_t >$. Denoting by σ_p^2 the variance in the fixed-time case with $< N_t >$ replaced by N_t, and using Eq. (3.24), a relationship between the variances in the two cases can be obtained:

$$\sigma^2 = \sigma_p^2 - \frac{1}{N_t} < C_{QL} >^2 \quad . \tag{3.40}$$

Since the second term in Eq. (3.40) is always non-negative, the fixed-count procedure always produces a smaller variance for corresponding numbers of photoevents. We intend to compare C_{QL} to a decision threshold to classify images, therefore we wish for C_{QL} to have the smallest variance attainable. We should, therefore, choose the fixed-photoevent-count configuration whenever possible. For the remainder of the chapter, we will continue to consider both experimental configurations because the fixed-time procedure can be advantageous for other reasons and because it may, in fact, be the only option available.

So far, in this chapter, we have considered the photon-limited image as a random signal and have developed its statistical properties. In the following sections, the theory of hypothesis testing for random signals is reviewed and shown to provide a useful approach to classifying quantum-limited images.

3.5 Statistical decision theory foundations

Consider the problem of choosing between two contradictory hypotheses on the basis of a set of noisy observations.[50] Let H_j denote the jth hypothesis and let d_i denote a decision in favor of hypothesis i. Further, let us assign a cost, c_{ij}, to making the decsion d_i when hypothesis H_j is true. Since the observations are not deterministic, a useful strategy for optimizing the decision-making process is to minimize the average cost of our decisions. The average (expected) cost per decision is given by

$$E\{c_{ij}\} = \sum_{j=1}^{2}\sum_{i=1}^{2} c_{ij}\, p(d_i, H_j) \quad , \tag{3.41}$$

or

$$E\{c_{ij}\} = \sum_{j=1}^{2}\sum_{i=1}^{2} c_{ij}\, p(H_j)\, p(d_i|H_j) \quad , \tag{3.42}$$

where the joint probability, $p(d_i, H_j)$ and conditional probability, $p(d_i|H_j)$, depend on our choice of decision strategy. Since the decisions d_1 and d_2 are complementary, the conditional probabilities are related by

$$\begin{aligned} p(d_1|H_1) &= 1 - p(d_2|H_1) \\ p(d_1|H_2) &= 1 - p(d_2|H_2) \end{aligned} \quad , \tag{3.43}$$

and substitution of Eq. (3.43) into (3.42) leads to

$$\begin{aligned} E\{c_{ij}\} = {} & c_{11}p(H_1) + c_{12}p(H_2) \\ & + (c_{21} - c_{11})p(H_1)p(d_2|H_1) \\ & + (c_{22} - c_{12})p(H_2)p(d_2|H_2) \quad . \end{aligned} \tag{3.44}$$

Consider the case in which our decision is to be based on an observation set ordered into a vector **z**. Consistent with the discussions of Chapter 1, consider the strategy of deciding in favor of hypothesis 2 when **z** lies in some region Z and in favor of hypothesis 1 when **z** lies in the complement of Z. In this scheme,

$$p(d_2|H_j) = \int_Z dz p(\mathbf{z}|H_j) \quad , \tag{3.45}$$

and the expected cost per decision becomes

$$E\{c_{ij}\} = c_{11}p(H_1) + c_{12}p(H_2) + \int_Z d\mathbf{z}[(c_{21} - c_{11})p(H_1)p(\mathbf{z}|H_1) + (c_{22} - c_{12})p(H_2)p(\mathbf{z}|H_2)] \quad . \tag{3.46}$$

To minimize the expected cost, the region Z must be chosen to include all points for which the integrand in Eq. (3.46) is negative. The region Z that satisfies this condition is the set of all points **z** for which

$$(c_{21} - c_{11})p(H_1)p(\mathbf{z}|H_1) + (c_{22} - c_{12})p(H_2)p(\mathbf{z}|H_2) < 0 \quad , \tag{3.47}$$

or, assuming $c_{ij} > c_{jj}$, for which,

$$\frac{p(\mathbf{z}|H_2)}{p(\mathbf{z}|H_1)} > \frac{(c_{21} - c_{11})p(H_1)}{(c_{12} - c_{22})p(H_2)} \quad . \tag{3.48}$$

The quantity on the left-hand-side of Eq. (3.48) is known as the likelihood ratio, $l(\mathbf{z})$. The decision rule for minimum average cost (Bayes risk) can be written in its final form

$$l(\mathbf{z}) \underset{d_1}{\overset{d_2}{\gtrless}} \frac{(c_{21} - c_{11})p(H_1)}{(c_{12} - c_{22})p(H_2)} \quad , \tag{3.49}$$

where

$$l(\mathbf{z}) = \frac{p(\mathbf{z}|H_2)}{p(\mathbf{z}|H_1)} \quad . \tag{3.50}$$

The decision rule notation of Eq. (3.49) denotes that we decide d_2 when the "greater than" condition holds and d_1 when "less than" is the case.

In applications for which correct decisions have no cost (c_{11}=c_{22}=0) and incorrect decisions are equally weighted (c_{12}=c_{21}), the Bayes risk decision rule [Eq. (3.48)] reduces to a criterion that minimizes the probability of error. The so-called probability-of-error decision rule is given by

$$l(\mathbf{z}) \underset{d_1}{\overset{d_2}{\gtrless}} \frac{p(H_1)}{p(H_2)} \quad . \tag{3.51}$$

If, further, the prior probabilities for the two classes are equal, the well-known maximum-likelihood criterion results:

$$l(\mathbf{z}) \underset{d_1}{\overset{d_2}{\gtrless}} 1 \quad . \tag{3.52}$$

The maximum-likelihood strategy is simply to select the hypothesis that is most likely consistent with the observation set. Equivalently, it directs us to choose the hypothesis for which the corresponding probability density function, evaluated at the measured value of $\mathbf{z}$, is largest.

3.6 Likelihood-ratio solutions for image classification

The essential element of each of the decision criteria described above is a comparison of the likelihood ratio to some threshold value. In this section, the precise form for the likelihood ratio is derived in terms of measurable quantities for various configurations of the quantum-limited image classification experiment.

Let us begin by reiterating the definition of the likelihood ratio given in Eq. (3.50):

$$l(\mathbf{z}) = \frac{p(\mathbf{z}|H_2)}{p(\mathbf{z}|H_1)} \quad . \tag{3.53}$$

In general, $\mathbf{z}$ represents a vector of noisy observations and H_1 and H_2 denote two competing hypotheses that we wish to evaluate. Applied to the problem at hand, H_j can be defined as the hypothesis that the image we wish to classify belongs to the jth image class. In the context of quantum-limited imaging, the photoevents detected within the pixels of a low-light-level image fill the role of the observation set in $\mathbf{z}$. Denoting by the vector $\mathbf{n}$ the list of pixel photoevent counts constituting the input image, the likelihood ratio for the quantum-limited image classification experiment becomes

$$l(\mathbf{n}) = \frac{p(\mathbf{n}|H_2)}{p(\mathbf{n}|H_1)} \quad . \tag{3.54}$$

The hypothesis that an image belongs to class j is a compound hypothesis suggesting that the image is one of the many members of class j. Hence the probabilities in Eq.

(3.54) can be written as a linear combination of the probabilities associated with each of the member images,

$$p(\mathbf{n}|H_j) = \sum_{k=1}^{M_j} p[H_k^{(j)}]p[\mathbf{n}|H_k^{(j)}] \quad . \tag{3.55}$$

In Eq. (3.55), $H_k^{(j)}$ denotes the hypothesis that the input image is the kth element of class j, $p[H_k^{(j)}]$ is the *a priori* probability for that image, and M_j is the number of images in class j. Substituting from Eq. (3.55) into Eq. (3.54) gives the general form for the likelihood ratio for quantum-limited image classification:

$$l(\mathbf{n}) = \frac{\sum_{k=1}^{M_2} p[H_k^{(2)}]p[\mathbf{n}|H_k^{(2)}]}{\sum_{k=1}^{M_1} p[H_k^{(1)}]p[\mathbf{n}|H_k^{(1)}]} \quad . \tag{3.56}$$

The conditional probabilities in Eq. (3.56) are determined by the intensity distribution in the image plane of the detection system and depend on the manner in which the image acquisition is conducted.

a) Fixed-integration-time experiment

If the quantum-limited imaging system is operated for a fixed integration time, then, under the conditions discussed in Section 3.2, the photoevent counts measured within its pixels can be approximated as independent, Poisson-distributed random variables with density function

$$p[\mathbf{n}|H_k^{(j)}] = \prod_{i=1}^{N} \frac{[s_{k,i}^{(j)}]^{n_i} \exp[-s_{k,i}^{(j)}]}{n_i!} \quad , \tag{3.57}$$

where N is the number of pixels in the image. The quantity $s_{k,i}^{(j)}$, representing the conditional mean photoevent count for pixel i, i.e.,

$$s_{k,i}^{(j)} = E\{n_i | H_k^{(j)}\} \quad , \tag{3.58}$$

can be computed according to Eq. (3.9). Substitution of Eq. (3.57) into Eq. (3.56) yields the solution for the likelihood ratio in a fixed-integration-time image classification experiment:

$$l(\mathbf{n}) = \frac{\sum_{k=1}^{M_2} p[H_k^{(2)}] \prod_{i=1}^{N} [s_{k,i}^{(2)}]^{n_i} \exp[-s_{k,i}^{(2)}]}{\sum_{k=1}^{M_1} p[H_k^{(1)}] \prod_{i=1}^{N} [s_{k,i}^{(1)}]^{n_i} \exp[-s_{k,i}^{(1)}]} \quad . \tag{3.59}$$

For computational purposes, it is useful to rearrange Eq. (3.59) to isolate the signal-dependent terms, thus producing the equivalent expression

$$l(\mathbf{n}) = \frac{\sum_{k=1}^{M_2} \left[\left\{ p[H_k^{(2)}] \prod_{i=1}^{N} \exp[-s_{k,i}^{(2)}] \right\} \prod_{i=1}^{N} [s_{k,i}^{(2)}]^{n_i} \right]}{\sum_{k=1}^{M_1} \left[\left\{ p[H_k^{(1)}] \prod_{i=1}^{N} \exp[-s_{k,i}^{(1)}] \right\} \prod_{i=1}^{N} [s_{k,i}^{(1)}]^{n_i} \right]} \quad . \tag{3.60}$$

b) Fixed-photoevent-count experiment

If, instead, the total number of photoevents composing the image is fixed, i.e., if the experiment is stopped after some number of photoevents, N_t, is counted, then according to Section 3.2, the pixel photoevent counts obey the multinomial distribution,

$$p(\mathbf{n} | H_k^{(j)}) = N_t! \prod_{i=1}^{N} \frac{[s_{k,i}^{(j)} / N_t]^{n_i}}{n_i!} \quad , \tag{3.61}$$

where N_t is the sum of the pixel photevent counts, n_i.

The resulting solution for the likelihood ratio is

$$l(\mathbf{n}) = \frac{\sum_{k=1}^{M_2} p(H_k^{(2)}) \prod_{i=1}^{N} [s_{k,i}^{(2)}]^{n_i}}{\sum_{k=1}^{M_1} p(H_k^{(1)}) \prod_{i=1}^{N} [s_{k,i}^{(1)}]^{n_i}} \quad . \tag{3.62}$$

Note that the fixed-integration-time and fixed-photoevent-count solutions are almost the same in form, differing only by the exponential terms that appear in the fixed-time solution.

In the absence of dead-time effects, the solutions presented above for cases (*a*) and (*b*) are exact. In applications for which execution time is not critical, or in which parallel processing capability can be brought to bear, these solutions can be used to directly calculate the log-likelihood ratio for image classification. Unfortunately, the required computation is more lengthy than application of the matched filter for every training image. The exact solution is useful, therefore, only if the log-likelihood ratio based on the training images can be used to estimate the log-likelihood ratio for images not included in the training set. This can be achieved by limiting the variation among the images within each class. This restriction limits the merging of classes for multiple-class sorting, but is not otherwise prohibitive.

Let us consider the practical implementation of the exact solutions given in Eqs. (3.60) and (3.62). We begin by noting that the terms preceding the products in these equations are constants that are independent of the input image. These terms can, therefore, be computed and stored in advance of system operation. At low light levels, computational efficiency can be enhanced by calculating the products of

powers of $s_{k,i}^{(j)}$ in Eqs. (3.60) and (3.62) as a series of multiplications. In that case, the powers of $s_{k,i}^{(j)}$ can be computed concurrently with the arrival of the photoevent coordinate information by maintaining an accumulated product for each training image (each term indexed by k in the summations). When a photoevent is detected in pixel i, each product is updated by multiplying it by its corresponding $s_{k,i}^{(j)}$. When the predetermined integration time of the experiment elapses or the prescribed number of photoevents has been detected, the products are combined with the signal-independent terms to compute the log-likelihood ratio according to Eq. (3.60) or Eq. (3.62).

The calculations required by the exact solutions are likely to be impractical for many applications, particularly if large training sets and multiple classes are involved. The alternative, in such cases, is to seek approximate solutions for the likelihood ratio that are more readily computed. Since we are well-equipped to perform inner product calculations, let us consider approximations to the above solutions that are linear in the photoevent counts n_i.

a) Least-square-error solution

One approach to developing a linear approximation to the likelihood ratio is to compute the function that best fits the exact solution in a mean-square-error sense. The derivation of this solution follows that of the Wiener filter.[51]

Any monotonic function of the likelihood ratio is an acceptable decision metric that affects only the decision thresholds defined in Sec. 3.5. For reasons that

will subsequently become apparent, the natural logarithm of the likelihood ratio, or log-likelihood ratio, often takes the place of $l(\mathbf{n})$ in the decision-making process.

Considering the log-likelihood ratio as our new decision metric, the linear approximation that we seek has the form

$$\ln\{l(\mathbf{n})\} \equiv \sum_{i=1}^{N} n_i h_i \quad , \tag{3.63}$$

where, again, N is the number of pixels in the input image and n_i is the photoevent count for pixel i. The problem of finding the optimal linear solution lies in determining the best choice of $\mathbf{h}$. To begin, we define the mean-square-error criterion as the mean-square deviation of the linear solution from the actual solution as a function of the discriminant vector $\mathbf{h}$, i.e.,

$$\varepsilon(\mathbf{h}) = E[\{L(\mathbf{n}) - <\mathbf{h},\mathbf{n}>\}^2] \quad , \tag{3.64}$$

where $<\mathbf{h},\mathbf{n}>$ is shorthand notation for the inner product of Eq. (3.63) and $L(\mathbf{n})$ denotes the log-likelihood ratio $\ln\{l(\mathbf{n})\}$.

For $\mathbf{h}$ to provide the minimum value of ε, it must be the case that any other choice produces a larger value. Hence, a necessary condition for $\mathbf{h}$ to provide a minimum is

$$\varepsilon(\mathbf{h} + \mu\hat{\mathbf{h}}) \geq \varepsilon(\mathbf{h}) \quad , \tag{3.65}$$

for all real constants μ and all real vectors $\hat{\mathbf{h}}$. Equivalently we can write

$$\varepsilon(\mathbf{h} + \mu\hat{\mathbf{h}}) - \varepsilon(\mathbf{h}) \geq 0 \quad . \tag{3.66}$$

Substituting from Eq. (3.64) into Eq. (3.66) and expanding the square in Eq. (3.64) yields the condition

$$2\mu\{E[<\mathbf{h},\mathbf{n}><\hat{\mathbf{h}},\mathbf{n}>]-E[<\hat{\mathbf{h}},L(\mathbf{n})\mathbf{n}>]\}+\mu^2E[<\hat{\mathbf{h}},\mathbf{n}>^2]\geq 0\,. \qquad (3.67)$$

Clearly, the quadratic μ-term is non-negative. If the quantity in braces in the linear μ-term is different from zero, then a value of μ exists that causes the left-hand-side of Eq. (3.67) to become negative, thus violating the condition. Hence, the quantity in braces must be exactly equal to zero, i.e.,

$$E[<\mathbf{h},\mathbf{n}><\hat{\mathbf{h}},\mathbf{n}>]-E[<\hat{\mathbf{h}},L(\mathbf{n})\mathbf{n}>]=0\quad . \qquad (3.68)$$

Equation (3.68) can be rewritten as

$$E[\sum_{i=1}^{N}\hat{h}_i\sum_{i'=1}^{N}h_{i'}n_in_{i'}-\sum_{i=1}^{N}\hat{h}_iL(\mathbf{n})n_i]=0\quad , \qquad (3.69)$$

or

$$\sum_{i=1}^{N}\hat{h}_i\left\{\sum_{i'=1}^{N}h_{i'}E[n_in_{i'}]-E[L(\mathbf{n})n_i]\right\}=0\quad . \qquad (3.70)$$

Since Eq. (3.70) must hold for all $\hat{\mathbf{h}}$, it must be that the term in braces is equal to zero, i.e.,

$$\sum_{i'=1}^{N}h_{i'}E[n_in_{i'}]-E[L(\mathbf{n})n_i]=0\quad , \qquad (3.71)$$

or,

$$\sum_{i'=1}^{N}h_{i'}E[n_in_{i'}]=E[L(\mathbf{n})n_i]\quad . \qquad (3.72)$$

If we define a matrix **R**, having elements

$$R_{ii'}=E[n_in_{i'}]\quad , \qquad (3.73)$$

and a column vector **t**, having components

$$t_i = E\{L(\mathbf{n})n_i\} \quad , \tag{3.74}$$

then we can rewrite Eq. (3.72) as a matrix equation,

$$\mathbf{Rh} = \mathbf{t} \quad . \tag{3.75}$$

Solving Eq. (3.75), we obtain a condition that a vector **h** must satisfy to provide a least-square-error solution for the log-likelihood ratio,

$$\mathbf{h} = \mathbf{R}^{-1}\mathbf{t} \quad . \tag{3.76}$$

Unfortunately, direct computation of **h** from Eq. (3.76) is impractical for tasks involving large images. The matrix **R** has dimension N^2 where N is the number of pixels in the image. Images sampled on a 64×64 array, for example, produce a matrix **R** containing 4096×4096 elements. Furthermore, calculation of the elements of **R** is impractical except by estimation from realizations of **n**, either actual or simulated. It appears, therefore, that the least-square-error approach is too cumbersome to be widely applicable. We turn, therefore, to a less accurate, but far more practical approach.

b) Small-intraclass-variation approximation

Solutions for the log-likelihood ratio, of the linear form that we seek, can be obtained immediately from Eqs. (3.60) and (3.62) by assuming that the images within each class can be approximated by their corresponding class mean. If we take each of the training images to be approximately equal to the mean image of the class to which it belongs, i.e., if

$$s_{k,i}^{(j)} \cong m_i^{(j)} \quad , \tag{3.77}$$

where

$$m_i^{(j)} = \frac{1}{M_j}\sum_{k=1}^{M_j} s_{k,i}^{(j)} \quad , \tag{3.78}$$

then the likelihood ratio for the fixed-integration-time experiment becomes

$$l(\mathbf{n}) \cong \frac{\sum_{k=1}^{M_2} p[H_k^{(2)}]\prod_{i=1}^{N}[m_i^{(2)}]^{n_i}\exp[-m_i^{(2)}]}{\sum_{k=1}^{M_1} p[H_k^{(1)}]\prod_{i=1}^{N}[m_i^{(1)}]^{n_i}\exp[-m_i^{(1)}]} \quad . \tag{3.79}$$

The product terms in Eq. (3.79) are independent of k and can, therefore, be factored from the summation. Factoring, and using the fact that the prior probabilities sum to unity, i.e., that

$$\sum_{k=1}^{M_j} p[H_k^{(j)}] = 1 \quad , \tag{3.80}$$

Eq. (3.79) becomes

$$l(\mathbf{n}) \cong \frac{\prod_{i=1}^{N}[m_i^{(2)}]^{n_i}\exp[-m_i^{(2)}]}{\prod_{i=1}^{N}[m_i^{(1)}]^{n_i}\exp[-m_i^{(1)}]} \quad , \tag{3.81}$$

or

$$l(\mathbf{n}) \cong \prod_{i=1}^{N}\left[\frac{m_i^{(2)}}{m_i^{(1)}}\right]^{n_i}\exp[m_i^{(1)} - m_i^{(2)}] \quad . \tag{3.82}$$

Equation (3.82) can be simplified by taking its natural logarithm. The resulting quantity, known as the log-likelihood ratio, is given by

$$\ln\{l(\mathbf{n})\} \cong \sum_{i=1}^{N} n_i \ln\left[\frac{m_i^{(2)}}{m_i^{(1)}}\right] + \sum_{i=1}^{N}[m_i^{(1)} - m_i^{(2)}] \tag{3.83}$$

for the fixed-time experiment. The first term in Eq. (3.83) is simply the inner product between the image vector, **n**, and a discriminant vector **h**, the components of which are given by

$$h_i = \ln\left[\frac{m_i^{(2)}}{m_i^{(1)}}\right] \quad . \tag{3.84}$$

The discriminant vector **h** depends only on the training set images and can, therefore, be calculated in advance of system operation. The second term in Eq. (3.83) is independent of the input image and acts merely as a bias that must be included in the decision rule. If the mean images have the same total intensity, then this term vanishes.

Under the assumption of small intraclass variations, the log-likelihood ratio for the fixed-photoevent-count experiment (case (*b*)) can be derived in a similar fashion and differs only in that the bias term does not appear. The approximate fixed-photoevent-count solution for the log-likelihood ratio is

$$\ln\{l(\mathbf{n})\} \cong \sum_{i=1}^{N} n_i \ln\left[\frac{m_i^{(2)}}{m_i^{(1)}}\right] \quad . \tag{3.85}$$

3.6.1 Experimental results

The experiments described in this section were designed to: 1) evaluate the relative utility of the decision-theoretical approach when applied to actual image classification problems, 2) to demonstrate the approach in a working system, and 3) to test the validity of the small-intraclass-variation approximation introduced in Section 3.1.

To consider the first question, two pairs of image classes were assembled: printed characters (F and R) and tools (hammers and pliers). The character images [Fig. 3.4(a)], as in the experiments described in Chapter 2, were separated into two groups: the training set comprised seven fonts; five fonts were reserved as test examples. Likewise, each class of tool images [Fig. 3.4(b)] was divided into a training set containing ten images and a test set composed of four images. Each of the images in the experiment was a 64×64 pixel array and each was acquired by forming a histogram of 10^6 photoevent counts reported by the imaging photon-counting detection system described in Chapter 1.[31,32] To begin, the training images were used to construct discriminant vectors for implementation of the Fukunaga-Koontz transform, the average filter method, and the approximate log-likelihood solution described above.

The Fukunaga-Koontz basis vectors, generated in the manner described in Section 2.6, were ordered into a two-dimensional image format, and used in place of the discriminant vector $\mathbf{h}$ in Eq. (3.14).

(a)

(b)

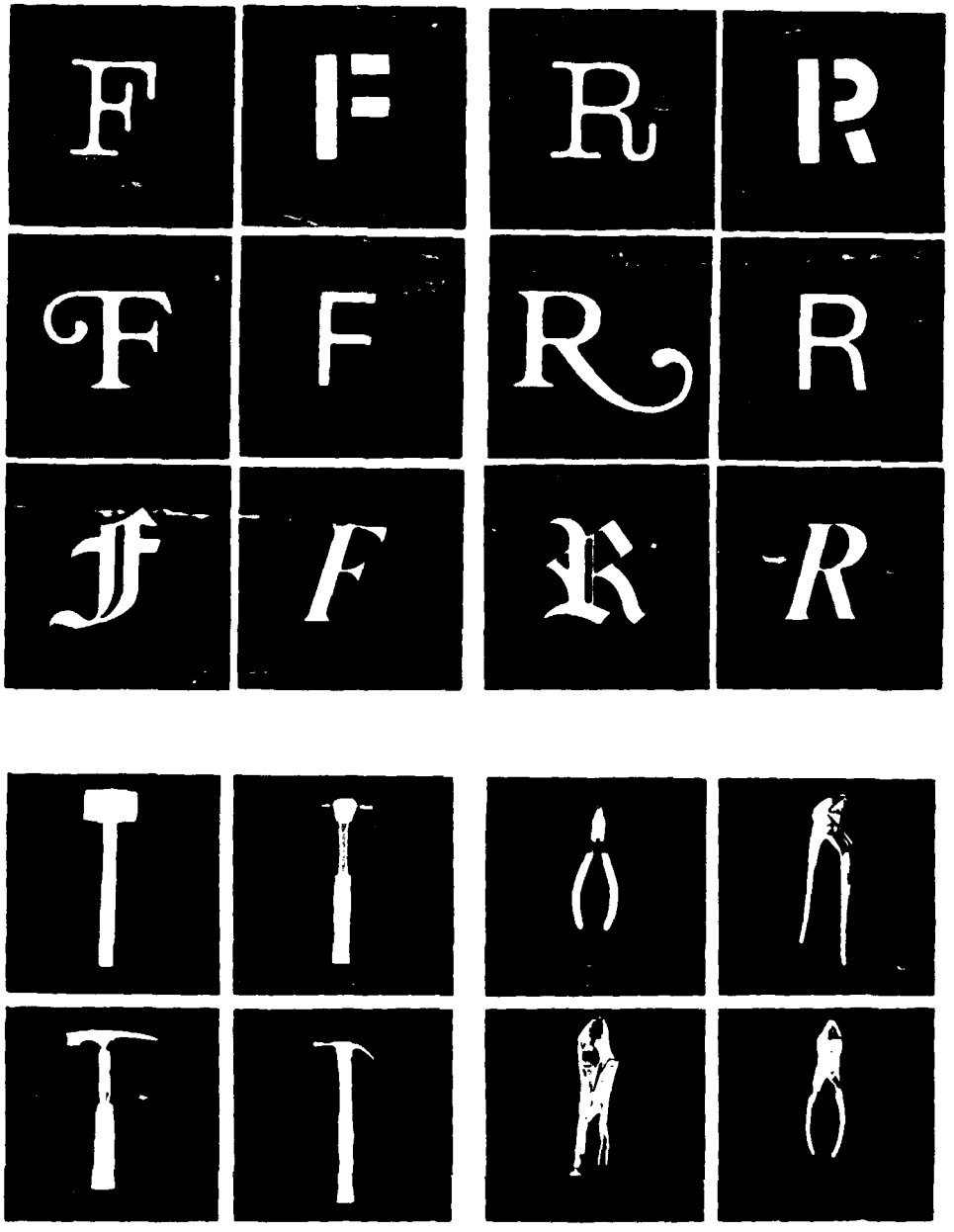


Fig. 3.4 Examples of images used in low-light-level experiments. (a) Characters: seven fonts were used as prototypes; five were reserved as test examples. (b) Tools: ten images from each class were used for training; four were used as test examples.

(a)

(b)

(c)

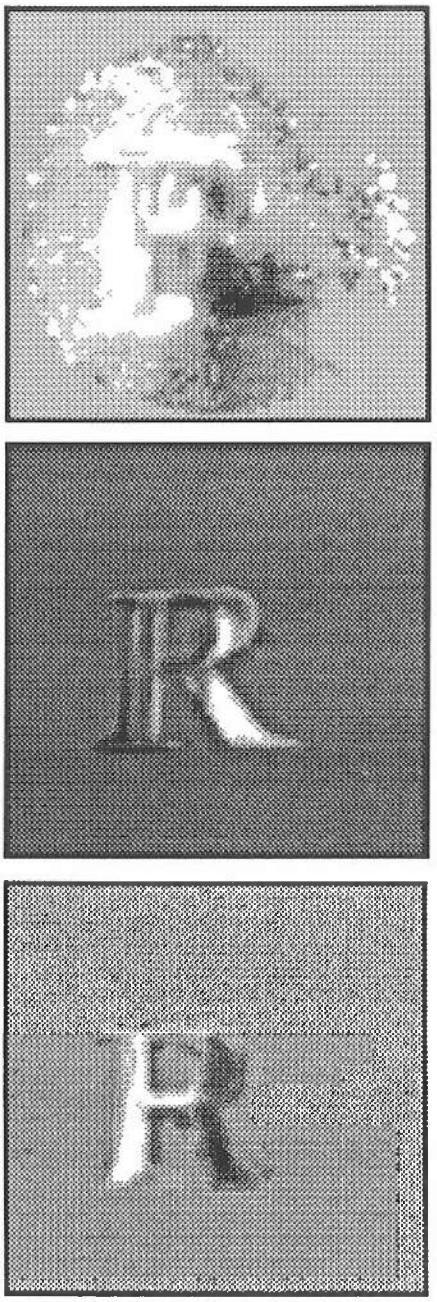


Fig. 3.5 Two-dimensional representations of discriminant vectors for character experiment. (a) Log-likelihood; (b) Difference of means; (c) Fukunaga-Koontz.

The log-likelihood discriminant vector was computed according to Eq. (3.85) from the training images for each two-class experiment. Care must be taken in calculating this vector since a singularity arises when all of the training set elements representing one of the classes have a zero in intensity at the same pixel. In theory, this singularity is part of the method; in practice, it creates undesirable sensitivity to noise. Consider, for example, the problem of distinguishing the letters F and R. The pixels constituting the tail of the letter R are exposed to zero intensity when an image of an F is presented to the system. If the set of R's is denoted as class 1 and the set of F's is labelled as class 2, then the log-likelihood discriminant vector [Eq. (3.84)] will take on the value $-\infty$ in those pixels containing the R's tail. The singularity indicates that the detection of a photon at that point is a zero-probability event if the input image is an F. If a photon is, in fact, detected there, the log-likelihood ratio takes on the value $-\infty$ and the decision for R is made immediately. In practice, of course, this is a dangerous procedure, since dark current, stray light, and outlier images can cause wild fluctuations in the value of the likelihood ratio. To suppress this effect, a bias can be added to each image in proportion to the level of additive noise expected in the experiment. In the experiments reported in this section, a uniform bias corresponding to a signal-to-noise ratio of 250 was added. Experience has shown that the results are not particularly sensitive to the amount of bias.

The Fukunaga-Koontz, difference-of-means, and decision theory approaches were compared using a figure of merit constructed for the low-light-level experiment (the discriminant vector for each appears in Fig. 3.5). The practical advantage of the low-light-level inner product technique is determined by the number of photoevents

required to make the discrimination decision. It is reasonable, therefore, to use the number of photoevents necessary to produce a particular error rate as a measure of performance. In these experiments, an error rate of one part in 10^4 was chosen as a point of reference. The number of photoevents required to produce this error rate was determined using the statistical considerations described in Section 3.4. The results of this experiment are tabulated in Table 3.1. Note that the performance of the algorithms has the same ranking for both sets of image classes.

Table 3.1. Number of Detected Photoevents Required to Achieve Probability of Error of 10^{-4}

Discriminant vector	Characters	Tools
Log-likelihood	146	43
Difference of means	246	91
Fukunaga-Koontz	965	136

The detector described in Section 3.1 operates at a rate on the order of 10^5 counts/sec. It can, therefore, produce classification decisions in 1.46 msec for the characters; in 0.43 msec for the tools. The current generation of detectors operate at count rates on the order of 10^6 counts/sec. The speed of a system based on such a detector would be limited by the electronics. If this could be overcome, the system could, in principle, produce classification decisions in 146 μsec for the characters; in 43 μsec for the tools.

The second objective pursued in the experiments was to demonstrate the feasibility of applying these techniques in a working system. Toward this end, laboratory experiments were performed in which the character images were classified using the photon-counting imaging system described in Section 3.1. The resolution

of the device has been estimated at 400×400,[25] however to simplify the experiments, the coordinates were digitized to 6-bit accuracy to produce a 64×64 pixel array of possible photoevent locations. The digital coordinates were passed to an HP1000 computer that calculated the inner product estimates according to the procedure described in Section 3.3. The computing power of the HP1000 was useful for algorithm development and research purposes. In operation, however, a much smaller machine would be adequate. For practical applications, a hardware implementation in which the photoevent coordinates are distributed to parallel, inner product computing units would be ideal.

In the experiments, the character images, recorded as transparencies, were imaged using incoherent illumination provided by an incandescent source. Appropriate neutral density was inserted in the camera to produce a count rate of 4×10^4 counts/sec. In the experiments, the count rate due to dark current (approximately 60 counts/sec) was negligible compared to that due to signal and hence was not considered in the analysis of the results.

Estimates of the inner products between an input image and the two best Fukunaga-Koontz basis vectors were calculated using 2000 detected photoevents. This experiment was performed 1000 times for each of the 24 character images. The theoretical mean values of these inner products are plotted in Fig. 3.6(a). The means correspond to the results that would be obtained in a deterministic evaluation of the inner product (e.g., in an optical correlator). In Fig. 3.6(b) the experimental realizations of the inner product are plotted. Each point represents the result of one classification. Each surrounding ellipse, representing the theoretical equal-probability

contour enclosing 99.9% of the integrated probability, was computed on the basis of the results of Section 3.4. Note that the theoretical contours and experimental realizations are in excellent agreement.

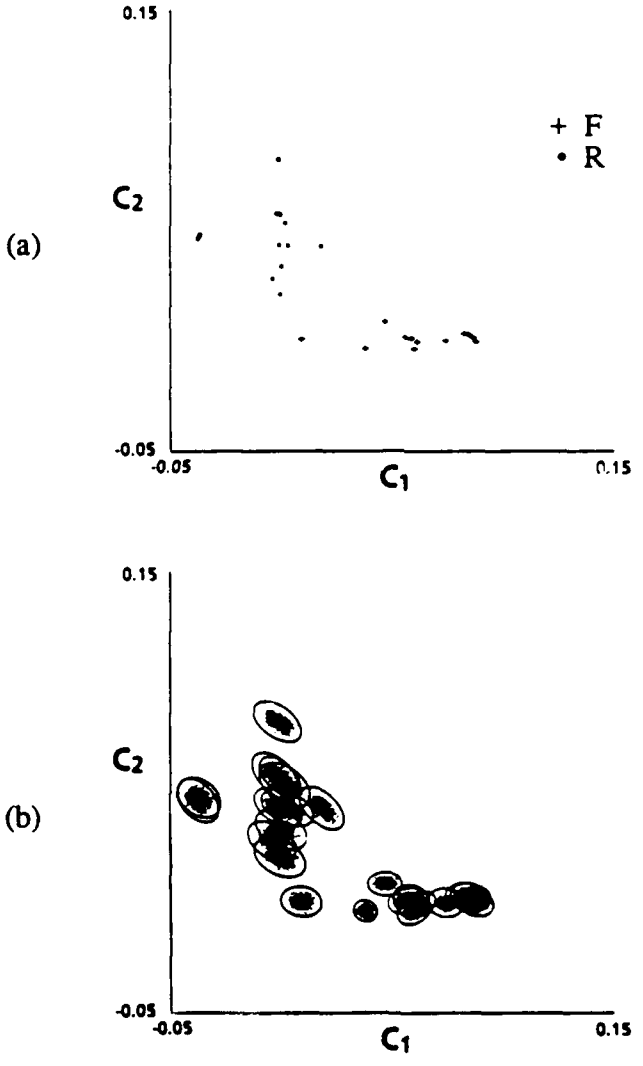


Fig. 3.6 Quantum-limited inner products using Fukunaga-Koontz disciminant vectors. (a) Experimental mean values; (b) experimental realizations and theoretical probability contours enclosing 99.9% of total prob ıbility.

In the Fukunaga-Koontz algorithm, classification of an input image is achieved by calculating the pair of inner products (C_1 and C_2) and comparing it with the linear decision boundary T. The boundary shown in Fig. 3.6 was chosen to minimize the probability of error. Theory predicts that the probability of error for the data appearing in Fig. 3.6(b) was approximately 2.4×10^{-7}. Of the 2.4×10^4 classifications represented in Fig. 3.6(b), there were no errors.

The conclusions of this experiment are that the use of the photon-limited imaging system for image classification operates in practice as predicted by theory. The advantages discussed in Section 3.3 are, in fact, realized in the working system.

The final objective pursued in the experiments was to determine whether the small-intraclass-variation assumption introduced in Section 3.6 was reasonable in practice. To accomplish this, the exact and approximate forms for the log-likelihood ratio were calculated using simulated photon-limited images produced by a random-number generator. The images were generated to simulate a fixed-integration-time scheme in which the average number of photoevents composing each image was 146 (see Table 3.1). Both the exact and approximate forms for the log-likelihood ratio were calculated for 1000 realizations of each of the 24 character images. Histograms of the results [Figs. 3.7(a) and 3.7(b)] illustrate that the approximate solution exhibits the correct behavior, but is consistently smaller in magnitude than the exact solution. The average ratio of the approximate to the exact solution is about 0.78. The performance obtained using the decision theory approach appears to be unharmed by the disagreement between the approximate and exact solutions due to its monotonic nature. By both the approximate and exact calculations, only two errors were made in 24,000

classifications, corresponding to an overall error rate of 8.3×10^{-5}, slightly below that predicted by theory.

Although the approximate solution for the log-likelihood ratio appears to be somewhat imprecise, it provides an excellent practical solution to the problem, comparing very well with other techniques.

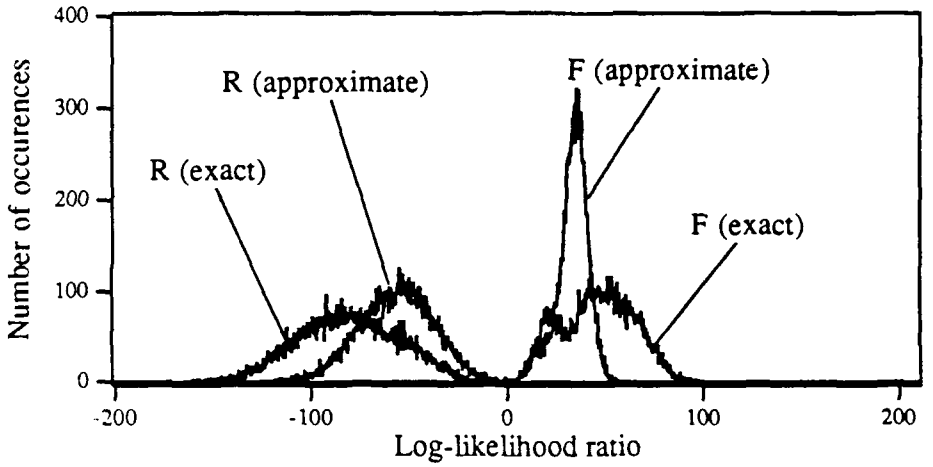


Fig. 3.7 Histograms of log-likelihood ratio values for character experiment produced by computer simulation. "Approximate" and "exact" indicate solution employed.

3.7 Rotation-invariant image classification

The statistical decision theory method described in the previous sections takes no account of geometrical variations of the image. The focus of this section is to demonstrate that techniques from the pattern recognition literature can be combined with the statistical decision theory formulation to introduce invariance to geometrical distortions. The use of the circular-harmonic components of images for rotation-invariant matched filtering has been described by Hsu and Arsenault[52] and by Isberg and Morris.[53] Here, we consider the circular-harmonic expansion as a means to eliminate sensitivity of the quantum-limited classification system to in-plane rotation of the input image.

Consider an experiment in which we wish to analyze an input image, represented as a two-dimensional spatial function g, by forming the inner product between g and some reference function, f. If the reference function is represented in polar coordinates, then it can be described in terms of an orthogonal expansion in the following way:

$$f(r,\theta) = \sum_{m=-\infty}^{\infty} F_m(r)\exp(im\theta) \quad , \tag{3.86}$$

where

$$F_m(r) = \tfrac{1}{2\pi}\int_0^{2\pi} f(r,\theta)\exp(-im\theta)d\theta \quad . \tag{3.87}$$

The mth term of the expansion in Eq. (3.86) is known as the mth circular-harmonic component of f and is denoted by $F_m(r,\theta)$. Consider the result of forming the inner product between a rotated input image and the complex conjugate of the mth

harmonic. If the image, represented in polar coordinates as $g(r,\theta)$, is rotated by an angle α, then the inner product between g and the conjugate of the mth harmonic is given by

$$C(\alpha) = \int_0^{2\pi} d\theta \int_0^{\infty} dr\, r\, F_m^*(r,\theta) g(r,\theta+\alpha) \quad . \tag{3.88}$$

Expanding the functions f and g and using the orthogonality of the circular harmonics, Eq. (3.88) becomes

$$C(\alpha) = 2\pi e^{im\alpha} \int_0^{\infty} dr\, r\, F_m^*(r) G_m(r) \quad . \tag{3.89}$$

Taking the norm of the inner product, we obtain

$$|C| = 2\pi \left| \int_0^{\infty} dr\, r\, F_m^*(r) G_m(r) \right| \quad . \tag{3.90}$$

The result is simply the norm of the inner product between the mth components of the image, g, and the reference function, f. The quantity $|C|$ contains information concerning the full inner product between f and g, however the dependence on the rotation angle α has been completely eliminated.

This property can be exploited for image classification by placing a discriminant vector $\mathbf{h}$ in the role of the function f. The specifics of the proposed experiment are as follows. A discriminant vector, computed for the image classes of interest, is arranged into a discrete two-dimensional image format and decomposed in its circular-harmonics. The real and imaginary parts of a subset of these harmonics are then installed in computer memory to ready the system for operation. When an image to be classified is presented to the system, the quantum-limited inner product between that image and each harmonic is computed in the manner described in Sec.

3.3. A linear combination of the moduli of the resulting inner products is compared to some predetermined decision threshold to complete the classification. Since the inner product moduli are rotation-invariant, the procedure is unaffected by changes in orientation of the input image.

To understand the behavior of the system when appled to photon-limited images, it is necessary to determine the effect of the stochastic nature of the input image on the inner product moduli and their linear combinations. This is particularly important for system design, in particular for selection of harmonics and decision boundaries and for evaluation of the number of photoevents needed to achieve the desired reliability level. The following is a summary of the essential statistical properties of the central quantities (a complete development of complex correlation statistics can be found in Ref. 54).

The real and imaginary parts (C' and C'') of the complex inner product between a discrete, photon-limited input image and one circular harmonic each obey a Gaussian distribution (see Section 3.4). The distribution for the real part is given by

$$p(C') = \frac{1}{\sigma'\sqrt{2\pi}} \exp\left[-\frac{(C' - < C' >)}{2\sigma'^2}\right] \quad , \tag{3.91}$$

where, for the fixed-photoevent-count experiment,

$$< C' >= \left(\frac{N_t}{\sum_{i=1}^{N} x_i}\right) \sum_{i=1}^{N} x_i h_i' \quad , \tag{3.92}$$

and

$$\sigma'^2 = N_t \frac{\sum_{i=1}^{N} x_i h_i'^2}{\sum_{i=1}^{N} x_i} - N_t \left(\frac{\sum_{i=1}^{N} x_i h_i'}{\sum_{i=1}^{N} x_i} \right)^2 \quad . \tag{3.93}$$

In Eqs. (3.92) and (3.93), h_i' represents the real part of one circular-harmonic component of the discriminant vector, sampled on the pixel grid, and ordered into the image vector format. As before, **x** denotes the classical-intensity image underlying the photon-limited input image **n**. The distribution for the imaginary part of the inner product has exactly the same form as that describing the real part and is obtained by replacing the primes in Eqs. (3.91)-(3.93) with double primes.

The joint density for the real and imaginary parts of the inner product is simply a general bivariate normal distribution,

$$p(C',C'') = \frac{1}{2\pi\sigma'\sigma''\sqrt{1-\rho^2}} \exp\left\{ \frac{1}{2(1-\rho^2)} \left[\frac{(C' - < C' >)^2}{\sigma'^2} - \frac{2\rho(C' - < C' >)(C'' - < C'' >)}{\sigma'\sigma''} + \frac{(C'' - < C'' >)^2}{\sigma''^2} \right] \right\} , \tag{3.94}$$

where the correlation coefficient, ρ, is defined as

$$\rho = \left\langle \frac{(C' - < C' >)(C'' - < C'' >)}{\sigma'\sigma''} \right\rangle \quad . \tag{3.95}$$

Because of its rotation-invariant properties, the modulus of the inner product, $|C|$, is the quantity of interest in this experiment. The density function for $|C|$ is derived by transformation of variables in Eq. (3.94) according to

$$C' = |C| \cos\gamma \, , \tag{3.96}$$

and

$$C'' = |C| \sin\gamma \,. \tag{3.97}$$

Using

$$|C| = \sqrt{C'^2 + C''^2} \quad , \tag{3.98}$$

we obtain:

$$p(|C|) = \frac{1}{2\pi\sigma'\sigma''\sqrt{1-\rho^2}} \int_0^{2\pi} d\gamma \exp\left\{ \frac{1}{2(1-\rho^2)} \left[\frac{(|C|\cos\gamma - < C' >)^2}{\sigma'^2} - \frac{2\rho(|C|\cos\gamma - < C' >)(|C|\sin\gamma - < C'' >)}{\sigma'\sigma''} + \frac{(|C|\sin\gamma - < C'' >)^2}{\sigma''^2} \right] \right\} \,. \tag{3.99}$$

The procedure for the proposed experiment is to use two circular-harmonic components of the discriminant vector and to make the classification decision on the basis of a linear combination of the moduli of the resulting inner products. To describe this experiment, we must compute the density function for a linear combination of the two inner product moduli, $a|C_1|+b|C_2|$, which in turn must be derived from the joint density, $p(|C_1|,|C_2|)$.

In the most efficient computational procedure, both inner product moduli, $|C_1|$ and $|C_2|$, are calculated from a single realization of the photon-limited input image, **n**. Alternatively, however, we might imagine computing $|C_1|$ and $|C_2|$ from distinct realizations of **n**. In the former arrangement, the joint density function for the two inner product moduli is a complicated expression, the computation of which is unwieldy. In the latter scheme, on the other hand, the statistical analysis is quite simple. The inner product moduli are independent, their joint density, $p(|C_1|,|C_2|)$, becomes the product of the individual densities, $p(|C_1|)$ and $p(|C_2|)$, and the density function for the linear combination, $p(a|C_1|+b|C_2|)$, is simply a convolution of the

these densities. As will be shown in Section 3.7.1, the results obtained by the two procedures are almost identical. The proposed method, therefore, is to use the "same-realization" procedure in the operation of the system, but to utilize the "distinct-realization" construction to predict the results and to aid in system design.

3.7.1 Experimental results

The experiments described in this section† demonstrate the combined use of the statistical decision theory approach of Section 3.6 and the circular-harmonic expansion technique described above to achieve rotation-invariant classification of quantum-limited images. The results are compared with similarly obtained results based on the average filter.

In these experiments, the character and tool image sets illustrated in Sec. 3.6.1 were used to form the average filter (Section 1.2.1) and log-likelihood discriminant vector (Eq. (3.84)). It has been demonstrated that at least two circular-harmonic components are required to achieve reliable results in a matched filtering experiment,[52] although no systematic method has been developed for their selection. In these experiments, it was determined by trial and error that the 2nd and 4th circular-harmonic components of the discriminant vectors were effective for classifying the tool images; the 1st and 2nd harmonics were found useful for the characters. The real and imaginary parts of the harmonics used in the character experiments are shown in Fig. 3.8.

---

† The work described in this section was done in collaboration with T. Isberg.

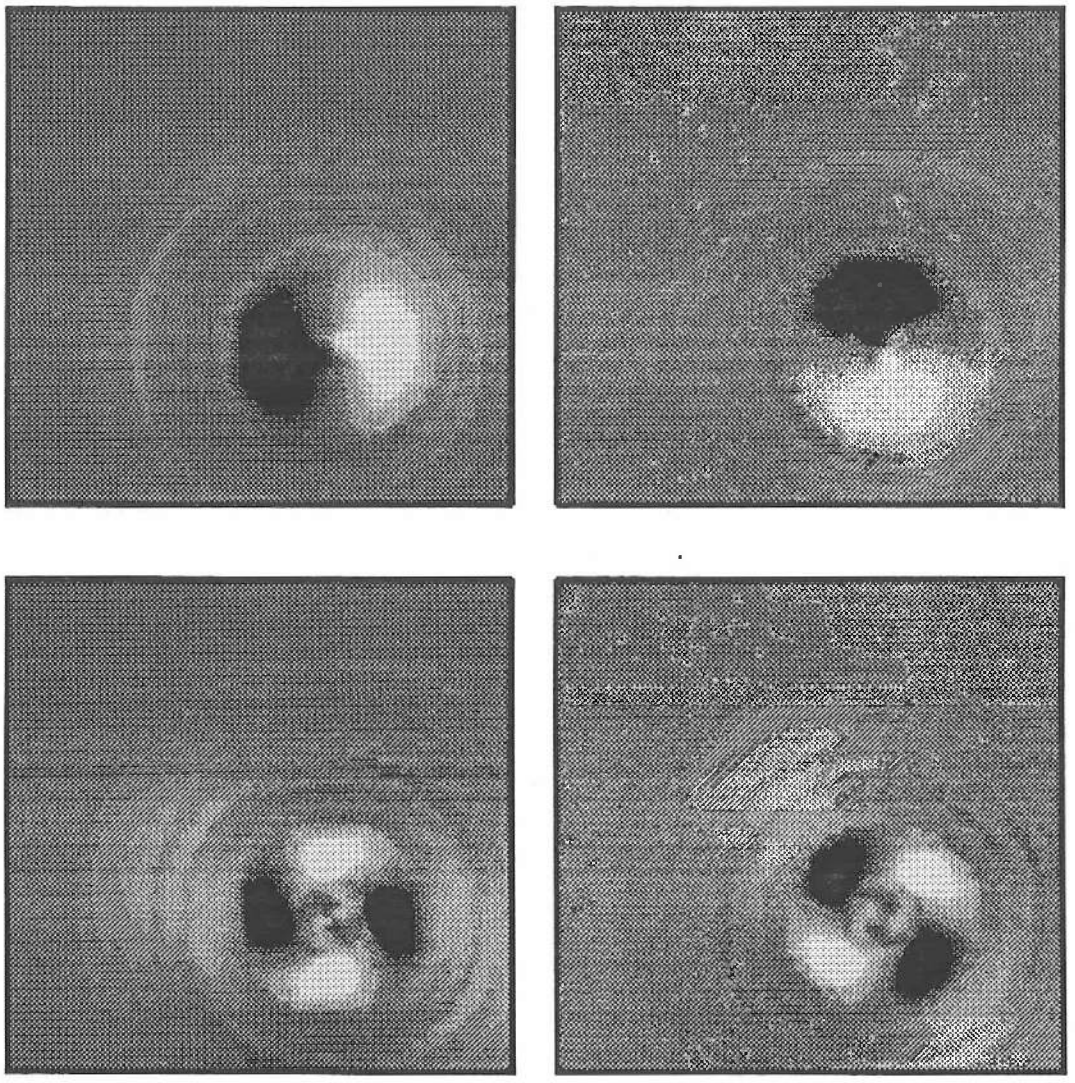


Fig. 3.8 Circular-harmonic components of log-likelihood discriminant vector for character experiment. Top row: real (left) and imaginary (right) parts of 1st harmonic; bottom row: real and imaginary parts of 2nd harmonic.

On the basis of the "distinct-realization" statistics and the proposed "same-realization" operational procedure, the number of photoevents required to reduce the probability of error to levels of 10^{-3} and 10^{-4} was computed for each image set and discriminant vector. The results of this analysis are tabulated in Table 3.2.

Table 3.2. Number of Detected Photoevents Required to Achieve Probabilities of Error (p_e) of 10^{-3} and 10^{-4}

Discriminant vector	Characters		Tools	
	$p_e = 10^{-3}$	$p_e = 10^{-4}$	$p_e = 10^{-3}$	$p_e = 10^{-4}$
Log-likelihood	256	450	590	930
Difference of means	338	542	1000	1620

To demonstrate feasibility in practice, experiments were performed using the imaging photon-counting setup described in Section 3.1. The real and imaginary parts of two circular harmonics of the log-likelihood discriminant vector were placed in computer memory. The inner product between an input character image and each part of each harmonic was formed using the photon-counting system illustrated in Section 3.1 by the method described in Section 3.3. The resulting real and imaginary inner products for each harmonic were combined to form the modulus of the complex inner product, which is invariant to rotation of the input image.

The experimental mean values of the inner product moduli for the 1st and 2nd circular-harmonic components of the log-likelihood discriminant vector are plotted against one another in Fig. 3.9. The linear decision boundary in Fig. 3.9 separates the class of F's from the class of R's. If the line is described by the equation,

$a|C_1|+b|C_2| = k_0$, then an input image, described by the point $(|C_1|,|C_2|)$, can be classified by comparing $a|C_1|+b|C_2|$ to k_0. If the result is greater than k_0, the image is taken to be an "F"; if the result is less than k_0 the image is perceived as an "R". Figures 3.10(a) and (b) each illustrate the results of 24,000 experimental realizations of the linear combination derived from the inner products between a rotated (90°) photon-limited character image (300 photoevents), and the 1st and 2nd circular-harmonic components of the log-likelihood discriminant vector. The results illustrated in Fig. 3.10(a) were obtained using the "same-realization" procedure; those shown in Fig. 310(b) by the "distinct-realization" scheme. Aside from statistical fluctuations, Figs. 3.10(a) and (b) are essentially identical.

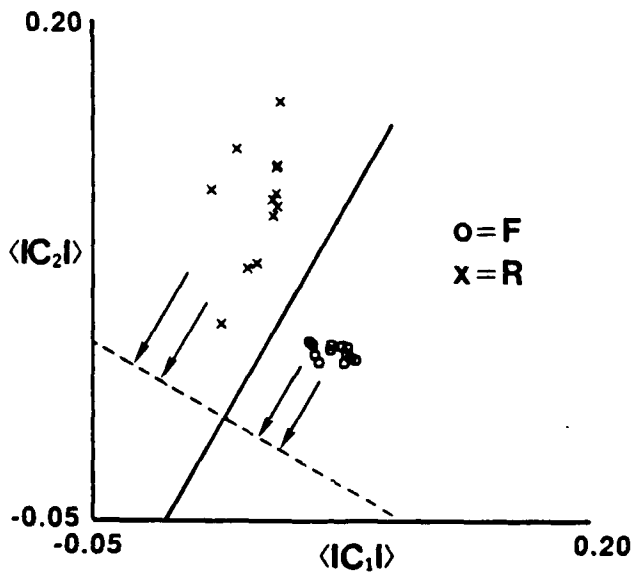


Figure 3.9 Mean values of inner products between character images and 1st and 2nd harmonics of log-likelihood discriminant vector.

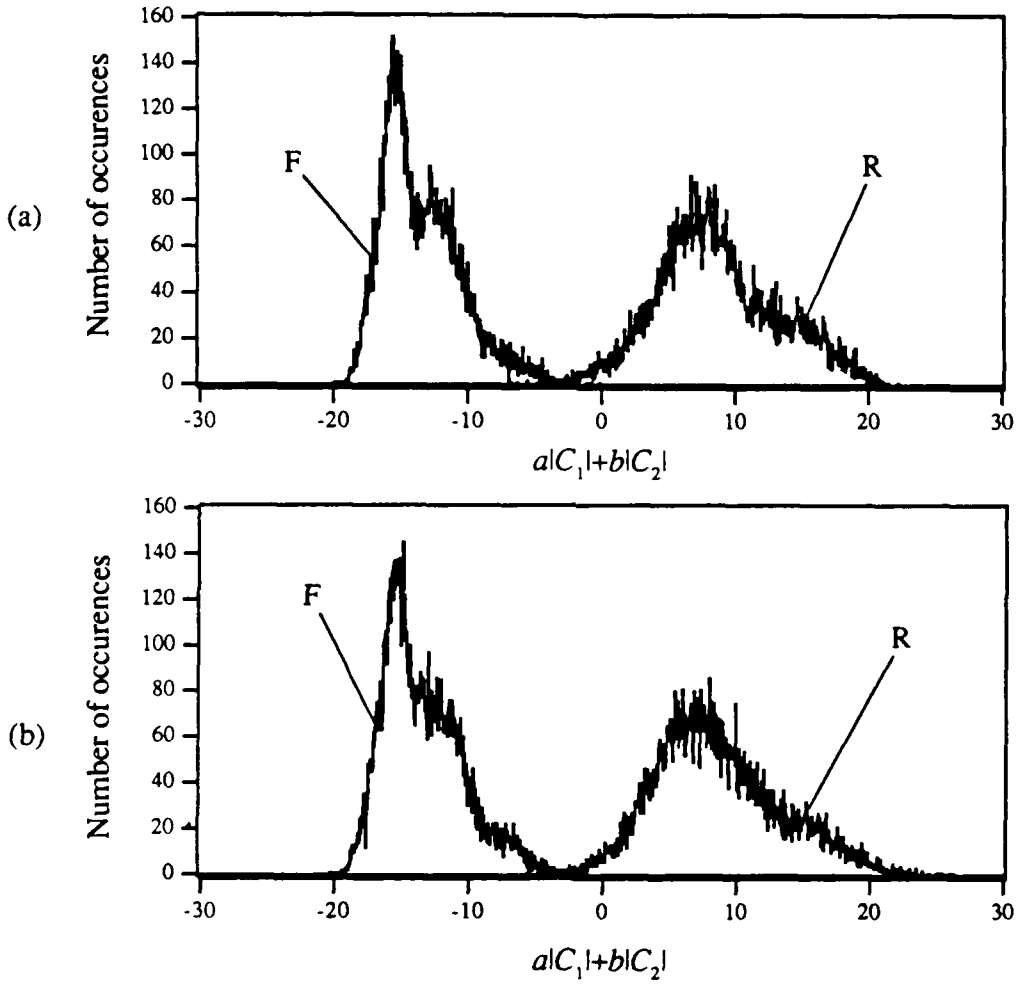


Figure 3.10 Histograms of linear combinations of inner product norms. (a) Same photoevents used for both harmonics; (b) Different photoevents used for 2nd harmonic. In each case $a = 0.874$ and $b = 0.485$.

Because only two circular-harmonic components are used in these experiments, some information contained in the full discriminant vector is lost. Hence, the number of photoevents required for reliable recognition is increased by the introduction of rotation-invariance. Fortunately, that number is still quite low. According to the results tabulated in Table 3.2, based on a count rate of 10^6 counts/sec and a prescribed error rate of 10^{-4}, the characters can be classified despite in-plane rotations in 0.93 msec; the tools in 0.45 msec.

3.8 Likelihood-ratio solutions applied to classical-intensity images

In the derivation of the approximate forms for the log-likelihood ratio [Eqs. (3.83) and (3.85)], no restriction was made on the total number of photons, N_t. These solutions can therefore be applied in the high-light-level limit of large N_t. In this section, computation of the log-likelihood ratio in a coherent optical image correlator is considered.

In attempting to adapt the solutions for the log-likelihood ratio to an optical correlator implementation, two issues must be addressed. First, the log-likelihood ratio can take on both positive and negative values, however we can only measure the modulus of the output of the optical system. Second, the optical correlator produces the full correlation function between the input image and the discriminant vector (arranged into a two-dimensional image format) but we are only interested in the inner product (the origin of the correlation function).

One solution to these problems can be obtained, in the fixed-photoevent-count case, by rewriting the log-likelihood discriminant vector as a difference of two vectors as follows:

$$\ln\left[\frac{m_i^{(2)}}{m_i^{(1)}}\right] = \ln\left[m_i^{(2)}\right] - \ln\left[m_i^{(1)}\right] \quad . \tag{3.100}$$

The log-likelihood ratio can then be approximated by a difference of inner products,

$$\ln\{l(\mathbf{n})\} \cong \sum_{i=1}^{N} n_i \ln\left[m_i^{(2)}\right] - \sum_{i=1}^{N} n_i \ln\left[m_i^{(1)}\right] \quad . \tag{3.101}$$

Provided that the $m_i^{(j)}$ are greater than one, each of the inner products in Eq. (3.101) will be positive and the modulus will not present a problem. This can always be achieved by first scaling all of the images by a constant multiplicative factor which cancels out in the difference operation.

Since the individual discriminant vectors in Eq. (3.101) resemble the images, the origin of the correlation function can, in some cases, be determined by locating the peak. While this procedure offers no guarantee, it can be useful in some instances.

3.8.1 Experimental results

The two-class character recognition problem considered in Section 3.6.1 was used to test the performance of the log-likelihood discriminant vector formulations at high light levels. Using the log-likelihood discriminant vector directly, the minimum separation and d' criteria introduced in Chapter 2 were computed. The results are shown in Tables 3.3 and 3.4, in which the results previously described in Chapter 2 are reiterated for comparison.

Table 3.3 Minimum Separation of Projections for Various Discriminant Vectors

Discriminant vector	Training images	Test images	Overall
Convex hull	0.391	0.284	0.284
Difference of means	0.131	0.308	0.131
Log-likelihood	0.124	0.204	0.124
Fukunaga-Koontz	0.062	0.071	0.062
F minus R	0.013	0.119	0.013
F	- 0.216	- 0.098	- 0.301

Table 3.4 Values of the d'-parameter for Various Discriminant Vectors

Discriminant vector	Training images	Test images	Overall
Convex hull	793.12	8.32	11.74
Log-likelihood	7.82	16.66	9.73
Difference of means	4.69	5.77	4.68
Fukunaga-Koontz	3.64	3.82	3.77
F minus R	2.91	5.00	3.16
F	1.49	1.87	0.28

To test the effectiveness of the two-discriminant-vector formulation, the output of an optical correlator was simulated using the Fast Fourier Transform (FFT) algorithm. Each input image was correlated with each discriminant vector ($\ln[m_i^{(1)}]$). The peak of the resulting correlation function was taken to be the inner product (defining the origin). These inner products are plotted against one another in Fig. 3.11. The boundary line indicates the maximum-likelihood decision threshold, $\ln\{l(\mathbf{n})\} = 0$. All but one of the characters are correctly classified by this decision

boundary; one falls on the line, resulting in no decision. All of the characters could have been correctly classified by adjusting the threshold in Fig. 3.11 upward, however this would no longer precisely constitute a maximum-likelihood strategy. In practice, this line can be determined from the prototype images and would be preferable despite the deviation from theory.

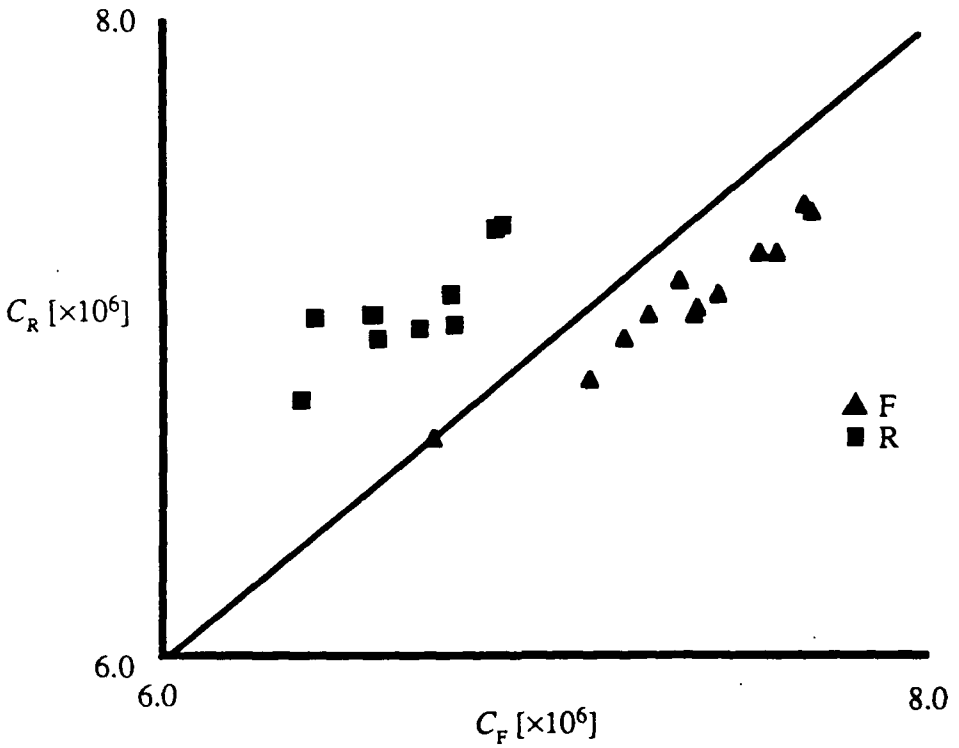


Figure 3.11 High-light-level inner products between separated log-likelihood discriminant vectors and character images.

References for Chapter 3

1. M. Lampton, Sci. Am. **245**, 62-71 (1981).
2. O. H. W. Siegmund, K. Coburn, and R. F. Malina, IEEE Trans. Nucl. Sci. **NS-32**, 433 (1985).
3. A. Boksenberg, in *ESO/CERN Conference of Auxiliary Instruments for Large Telescopes*, European Southern Observatory, Geneva, 295-316 (1972).
4. P. B. Boyce, Science **198**, 145-148 (1977)
5. A. Blazit, D. Bonneau, L. Koechlin, and A. Labeyrie, Astrophys. J. **214**, L75-L84 (1977).
6. E. D. Loh, "A search for a halo around NGC3877 with a charge coupled detector," Ph.D. dissertation. Princeton University, Princeton, NJ (1977).
7. ITT Electro-Optical Products Div., Application Note #E22. Fort Wayne, IN (1980).
8. C. R. Johnson and R. E. Blank, SPIE Semin. Proc. **290**, 102-108 (1981).
9. A. R. Jorden, P. D. Read, and I. G. van Breda, SPIE Semin. Proc. **331**, 368-375 (1982).
10. J. A. Tyson, J. Opt. Soc. Am. A **3**, 2131-2138 (1986).
11. E. Roberts, T. Stapinski, and A. Rodgers, J. Opt. Soc. Am. A **3**, 2146-2150 (1986).
12. E. M. Kellogg, S. S. Murray, and D. Bardas, IEEE Trans. Nucl. Sci. **NS-26**, 403-410 (1979).
13. J. G. Timothy, G. H. Mount, and R. L. Bybee, SPIE Semin. Proc. **183**, 169-181 (1979).

14. J. G. Timothy, G. H. Mount, and R. L. Bybee, IEEE Trans. Nucl. Sci. **NS-28**, 689-697 (1981).

15. J. G. Timothy, Opt. Eng. **24**, 1066-1071 (1985).

16. H. O. Anger, Instrum. Soc. Am. Trans. **5**,311-334 (1966).

17. C. Martin, P. Jelinsky, M. Lampton, R. F. Milina, and H. O. Anber, Rev. Sci. Instrum. **52**,1067-1074 (1981).

18. O. H. W. Siegmund, S. Clothier, J. Thorton, J. Lemem, R. Haper, I. Mason, and J. L. Culhane, IEEE Trans. Nucl. Sci. **NS-30**, 503-507 (1983).

19. O. H. W. Siegmund, R. F. Malina, K. Coburn, and D. Wertheimer, IEEE Trans. Nucl. Sci. **NS-31**, 776 (1984).

20. H. E. Schwartz and J. J. Lapington, IEEE Trans. Nucl. Sci. **NS-32**, 433-437 (1987).

21. O. H. W. Siegmund, M. Lampton, J. Bixler, S. Chakrabarti, J. Vallerga, S. Bowyer, and R. F. Malina, J. Opt. Soc. Am. A **3**, 2139-2145 (1986).

22. C. Papaliolios and L. Mertz, SPIE Semin. Proc. **331**, 360-364 (1982).

23. C. Papaliolios, P. Nisenson, and S. Ebstein, Appl. Opt. **24**, 287-292 (1985).

24. M. Lampton and C. W. Carlson, Rev. Sci. Instrum. **50**, 1093-1097 (1979).

25. C. Firmani, E. Ruiz, C. W. Carlson, M. Lampton, and F. Paresce, Rev. Sci. Instrum. **53**, 570-574 (1982).

26. D. Rees, I. McWhirter, P. A. Rounce, F. E. Barlow, and S. J. Kellock, J. Phys. E **13**, 763-770 (1980).

27. D. Rees, I. McWhirter, P. A. Rounce, and F. E. Barlow, J. Phys., E **14**, 229-233 (1981).

28. I. McWhirter, D. Rees, and A. H. Greenaway, J. Phys. E **15**, 145-150 (1982).

29. A. H. Greeenaway, A. Lyons, I. McWhirter, D. Rees, and A. Cochran, SPIE Semin. Proc. **331**, 365-367 (1982).

30. L. Mertz, T. D. Tarbell, and A. Title, Appl. Opt. **21**, 628-634 (1982).

31. ITT Electro-Optical Products Div., Tube and Sensor Laboratories, Fort Wayne, IN, detector Model F4146M.

32. Surface Science Laboratories, Inc., Mountain View, CA, Model 2401 position computer.

33. L. Mandel, Kinam Rev. Fis. Ser. C **5**, 213-232 (1963).

34. L. Mandel, E. C. G. Sudarshan, and E. Wolf, Proc. Phys. Soc. **84**, 435-444 (1964).

35. M. Bertolotti, in "Photon Correlation and Light Beating Spectroscopy," H. Z. Cummins and E. R. Pike, eds., p. 41, Plenum, New York, 1974.

36. B. Saleh, *Photoelectron Statistics*. Springer-Verlag, Berlin, 1978.

37. J. W. Goodman, *Statistical Optics*. Wiley, New York, 1985.

38. A. W. Lohmann, G. Weigelt, and B. Wirnitzer, Appl. Opt. **22**,4028-4037 (1983).

39. B. Wirnitzer, J. Opt. Soc. Am. A **2**, 14-21 (1985).

40. K.-H. Hofmann and G. Weigelt, Appl. Opt. **26**, 2011-2015 (1987).

41. H. Bartelt, A. W. Lohmann, and B. Wirnitzer, Appl. Opt. **23**, 3121-3129 (1984).

42. A. W. Lohmann, Optik **73**, 127-131 (1986).

43. J. C. Dainty and M. J. Northcott, Opt. Commun. **58**, 11-14 (1986).

44. D. Newman, A. A. D. Canas, and J. C. Dainty, "Space-time photon correlation using a 2-D photon event detector," Appl. Opt. **24**, 4210-4220 (1985).

45. A. Rose, *Vision: Human and Electronic*, (Plenum, New York, 1977).

46. W. O. Saxton and J. Frank, "Motif detetection in quantum noise-limited electron micrographs by cross-correlation," Ultramicroscopy **2**, 219-227 (1977).

47. G. M. Morris, "Scene matching using photon-limited images," J. Opt. Soc. Am. A **1**, 482 (1984).

48. G. M. Morris, "Image correlation at low light levels: a computer simulation," Appl. Opt. **23**, 3152-3159 (1984).

49. G. M. Morris, M. N. Wernick, and T. A. Isberg, "Image correlation at low light levels," Opt. Lett. **10**, 315-317 (1985).

40. J. L. Melsa and D. L. Cohn, *Decision and Estimation Theory*, (McGraw-Hill, New York, 1978).

51. W. B. Davenport, Jr. and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*, (McGraw Hill, New York, 1958).

52. Y.-N. Hsu and H. H. Arsenault, "Optical pattern recognition using the circular harmonic expansion," Appl. Opt. **21**, 4016-4019 (1982).

53. T. A. Isberg and G. M. Morris, "Rotation-invariant image recognition at low light levels," J. Opt. Soc. Am. A **3**, 964-971 (1986).

54. T. A. Isberg, Ph.D. Thesis, University of Rochester.

Chapter 4:
Concluding remarks

In this thesis, we have considered inner-product-based techniques for image classification. A new discriminant vector for classical-intensity image classification is described in Chapter 2. In this method, images are represented as vectors, the components of which are the pixel intensities. In the space defined by the pixel intensities, a set of points constitutes each image class. The discriminant vector introduced in Chapter 2 is the unit vector along the line segment connecting the points of closest approach of the convex hulls associated with these point sets. If the classes are linearly separable, then the proposed method produces no misclassifications. In addition, it is the linear disciminant for which the minimum separation of the inner product results is maximized.

In the experiments described in Chapter 2, the convex-hull discriminant vector approach was applied to two classification problems. In the first task, images of the letters F and R were identified despite significant font variations. Excellent results were obtained, including correct classification of images not incorporated in the training set. In the second task, eight letters in ten fonts were correctly classified using the theoretical minimum of three discriminant vectors.

In Chapter 3, a new approach to image classification is described. The images considered are composed of a very small number of photoevents measured by a position-sensitive photon-counting detection system. The inner product between the

photon-limited input image and a discriminant vector is shown to be an extremely simple operation, owing to the binary nature of the photoevents. Also demonstrated is that the ensemble average of the inner product between a photon-limited input image and a discriminant vector is directly proportional to the inner product that would be obtained using the underlying intensity image. Hence, any linear classifier can be implemented in the photon-counting system.

In addition, however, a method particularly suited to photon-limited images is developed in Chapter 3. An approximation is used to produce a linear discriminant that is used to implement statistical decision theory strategies. This method was shown, experimentally, to require fewer photoevents to achieve the same rate of misclassifications as the average filter and the Fukunaga-Koontz transform.

In Section 3.7, the circular-harmonic expansion is considered as a means to classifying images despite in-plane rotations. The norm of the complex inner product between a photon-limited input image and a circular-harmonic component of a discriminant vector is invariant to rotation of the image. Experiments described in Section 3.7.1 verify that the circular harmonics successfully classify both training and test images regardless of image orientation. Although rotation-invariant classification by this method requires a larger number of photoevents to attain the same reliability level, that number is still quite low.

Because of the small number of events needed to classify images, the quantum-limited image classification approach appears promising. It has obvious implications for applications such as night vision and low-dose electron microscopy, but may be

useful in high-light-level situations as well because of the speed and ease of operation of the system.

4.1 Future directions

Because of the complexity of the problem, research efforts in the area of image classification remain at an early stage of development. In their present state, algorithms like those described in this thesis can be quite successful when applied to constrained problems such as micrograph analysis, spectral pattern recognition, and printed-character recognition. However, complex problems, such as object recognition in cluttered, unpredictable environments remain unsolved. The primary issues that need to be investigated are intensity normalization, shift-invariance, and the advantages afforded by utilizing information from multiple sensors, perhaps using multiple algorithms.

The linear classifiers described throughout this thesis involve the comparison of an inner product to some threshold. As discussed in Chapter 1, it is usually the case that we wish to make no distinction between images $\mathbf{x}$ and $k\mathbf{x}$ ($k \in R^+$). Unfortunately, the inner product between $k\mathbf{x}$ and a discriminant vector $\mathbf{h}$ is k times the inner product between $\mathbf{x}$ and $\mathbf{h}$, therefore, unless the threshold is zero, the possibility of misclassifications presents a problem. The solution, of course, is to normalize the inner product (to factor out the effect of intensity variation).

An interesting and fortuitous feature of the photon-limited inner product system is that a kind of normalization is inherent in the process. It was shown in Chapter 3 that the ensemble average of the quantum-limited inner product is proportional to the high-light-level inner product, i.e.,

$$< C_{QL} >= \left(\frac{< N_t >}{\sum_{i=1}^{N} x_i} \right) \sum_{i=1}^{N} x_i h_i \quad . \tag{4.1}$$

Note that the denominator of the term in parentheses is the sum of the classical pixel intensities and acts to divide out the total intensity of the input image. The normalization factor that would be preferred, especially for the matched filter, is 1/||x||. Experiments have shown,[1] however, that the normalization provided by the photon-counting method is more than adequate.

Normalization in the coherent optical correlator is not so easily obtained and has not been widely considered in the literature, although it is clearly an important problem. A possible solution that warrants investigation in the future is revealed by considering the ideal system output.

Given a real input image $f(x,y)$ and an impulse response function $h(x,y)$, the desired correlation output is given by

$$g(x,y) = \frac{\iint_A dx'dy' f(x'+x,y'+y)h(x',y')}{\left[\iint_A dx'dy' f^2(x'+x,y'+y)\right]^{1/2}\left[\iint_A dx'dy' h^2(x',y')\right]^{1/2}} \quad , \tag{4.2}$$

where, for a matched filtering system, A denotes the support of the impulse response function. The numerator in Eq. (4.2) can be deduced from the standard optical correlator output. The term including the square-integral of h can be calculated in advance of system operation and is the same for any input image. The term that presents a difficulty is the one including the square-integral of f, and corresponds to ||x|| referred to in previous discussions.

A means to computing this normalization integral can be seen by rewriting the integral as

$$\iint_A dx'dy' f^2(x'+x,y'+y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dx'dy' f^2(x'+x,y'+y)w(x',y') \quad , \tag{4.3}$$

where w is a window function given by

$$w(x',y') = \begin{cases} 1 & ;(x',y') \in A \\ 0 & ;\text{otherwise} \end{cases} \quad . \tag{4.4}$$

In this form we see that the integral can be computed as a correlation between f^2 and w. This corresponds to the output of an incoherent (intensity-based) optical correlator with w applied as the impulse response.

In this scheme, computation of the normalized correlation function of Eq. (4.2) requires two side-by-side correlators. A coherent correlator computes the image correlation; an incoherent correlator computes the normalization factor. The two correlation planes are then measured and the quotient is computed electronically, yielding a single, normalized correlation plane.

A further area for future research lies in utilizing the shift-invariance of the optical correlator to implement the convex-hull approach described in Section 2. Like the methods described in Chapter 1, the convex-hull method is based upon an inner product operation. The inner product is simply the origin of a correlation function and can be computed in an optical correlator if the position of the object is known. The next step in investigation of the convex-hull approach is to give attention to the correlation functions produced by the discriminant vector and to adapt the method to create signatures in the correlation plane that identify object locations. Once found, these

points can then be examined using the direct convex-hull method described in this thesis.

Alternatively, or perhaps in addition, objects might be located using information gained from other sensors, e.g., infrared images, range images, or stereo images. Signatures within these images that may, themselves, be difficult to identify, might be used solely for indicating object positions which could then be scrutinized using a high-resolution optical image and the methods described in this thesis.

In closing, the solution of the most complex image classification problems may ultimately lie in the use of multiple sensors and the integration of more than one algorithm, perhaps combining the speed of optical detection methods with the flexibility of digital computation methods.

References for Chapter 4

1. Thomas Arthur Isberg, Ph.D. Thesis, University of Rochester, 1989.

Appendix A:
Rosen discriminant vector formulation

In a paper by Rosen (Ref. 3, Chapter 2), he shows that the linear separability of two classes is equivalent to the existence of a solution to the following quadratic programming problem:

Minimize with respect to $\mathbf{y}$ the function

$$\tfrac{1}{4}\sum_{i=1}^{N} y_i^2 \tag{A.1}$$

subject to

$$\begin{aligned} \mathbf{Q}_1^T\mathbf{y} &\geq \mathbf{e}_1 \\ -\mathbf{Q}_2^T\mathbf{y} &\geq \mathbf{e}_2 \quad . \end{aligned} \tag{A.2}$$

In expressions (A.1) and (A.2), $\mathbf{y}$ is an N-dimensional vector (in our context, N is the number of pixels), $\mathbf{e}_j$ is an M_j -dimensional vector of ones (M_j denotes the number of images in class j), and $\mathbf{Q}_j$ is an $(N+1) \times M_j$ matrix. Each column of $\mathbf{Q}_j$ contains an image vector belonging to class j followed by one element equal to -1. The solution for the discriminant vector is contained within the solution for $\mathbf{y}$ (it is the first N components).

The difficulty in solving the above optimization problem, when applied to image classification problems, lies primarily in the large number of dimensions in the choice variable $\mathbf{y}$. A classification problem involving images composed of 64×64 pixels leads to a 4096-dimensional optimization problem. In the new formulation, proposed in Chapter 2, the number of dimensions in the optimization is equal to the total number of prototype images, which is usually a great deal smaller than the number of pixels. Further, the proposed method involves fewer constraints than Rosen's method. The

new method requires two equality constraints and (M_1+M_2) non-negativity constraints, whereas the previous method involves (M_1+M_2) inequality constraints, which is equivalent to (M_1+M_2) equality constraints and an equal number of non-negativity constraints.

Appendix B:
Quadratic programming software

```
C-----------------------------------------------------------------------------------------
      SUBROUTINE QP(CONSTR, C, IDC, P, IDP, A, IRDA, ICDA, B, IDB, M, N,
     1     BASICX, IDBASICX, INLIST, IDINLIST, T, IRDT, ICDT, TSHOW, IER)
C-----------------------------------------------------------------------------------------
C
C     Solves quadratic programming problem:
C          max F(x) = p'x - x'C x/2
C            subject to   Ax=b  (m constraints)   (or Ax<=b)
C                         x>=0  (n dimensions)
C     (Can minimize by maximizing -F(x))
C
C     Variables:
C
C     CONSTR  = 1 for equality constraints
C             = 0 for inequality constraints
C
C     C,p,A,b are as defined. Dimension them at least as big as:
C     A=(mxn); x=(nx1); b=(mx1)
C     p=(nx1); C=(nxn)
C
C
C
C     ID[var]  =  dimension of 1-d array or square array [var] as
C                 specified in calling program
C     IRD[var] = row dimension of [var] as specified in calling program
C     ICD[var] = column dimension of same
C     m,n are defined above
C
C     BASICX(I) = list of m subscripts of basic x's in initial
C                   feasible solution (equality constraint problems only)
C                   Note: must have EXACTLY m!
C                   IDBASICX must be at least m
C
C     INLIST(I) = list of variables in the basis
C           INLIST(I)     = +j means ith variable is uj
C                         = -j means ith variable is xj
C                         =  0 means ith variable is vj
C     IDINLIST must be at least (m+n)
C
C     T(ROW,COL) = tableau
C     IRDT must be at least m+n
C     ICDT must be at least 2*(m+n)+1
C     TSHOW    = 1   ;display tableaux for diagnostic purposes
C              = 0   ;inhibit display
C     IER  = 1   ;error condition (# of iterations exceeded 2*(m+n))
C          = 0   ;no error
```

```
C     Declarations --

      implicit none
      integer constr,idc,idp,irda,icda,basicx(idbasicx),
     >        idb,m,n,i,j,ntot,row,col,row1,col1,idbasicx,Tshow,
     >        inlist(idinlist),bpairx,bpairu,nextin,nextout,sub,nbsub,
     >        minrow,us,rdT,cdT,iteration,irdT,icdT,idinlist,ier
      real   T(irdT,icdT),C(idC,idC),p(idp),A(irdA,icdA),b(idb),di,wi,diwi,min
      character*3 std
      character*45 outfile

      iteration = 0
      ntot = m + n
      ier=0

      do i = 1 , irdt
         do j = 1 , icdt
            t(i,j) = 0.0
         end do
      end do

C     Construct set-up tableau --

      do i = 1 , n
         inlist(i) = i              !count u's from 1
      end do
      do i = n+1 , ntot
         inlist(i) = -i             !count x's from n+1
      end do

C     Fill in values column --

      do row = 1 , n
         T(row,1) = -p(row)
      end do
      do row = n+1 , ntot
         T(row,1) = b(row-n)
      end do

C     Fill in -A' block --

      do row = 1 , n
         do col = n+2 , ntot+1
            col1 = col - (n+1)        !col1 = 1 , m
            T(row,col) = -A(col1,row)
         end do
      end do
```

```
C     Fill in A block --

      do row = n+1 , ntot
         do col = ntot+2 , ntot+n+1
            row1 = row - n                    !row1 = 1 , m
            col1 = col - (ntot+1)        !    col1 = 1 , n
            T(row,col) = A(row1,col1)
         end do
      end do


C     Fill in -C block --

      do row = 1 , n
         do col = ntot+2 , ntot+1+n        !1 is for values column
            col1 = col - (ntot+1)          !col1 = 1 , m
            T(row,col) = -C(row,col1)
         end do
      end do

C     Fill in I blocks --

      do row = 1 , n
         do col = 2 , n+1
            col1 = col - 1                 !col1 = 1 , n
            if (row.eq.col1) T(row,col) = 1
         end do
      end do
      do row = n+1 , ntot
         do col = (ntot+1)+(n+1) , 2*ntot+1
            row1 = row - n
            col1 = col - (ntot+n+1)
            if(row1.eq.col1) T(row,col) = 1
         end do
      end do

      if (Tshow.eq.1) call disp(T,irdt,icdt,ntot,2*ntot+1,inlist)
```

```
C     Are the constraints equality constraints?

      if (constr.eq.1) then

C     If so, form initial tableau by eliminating v,y from basis --

      do i = 1 , m                    !Replace u1's with v's
         nextout = basicx(i)
         nextin = i+n
         call update
     >     (T,irdT,icdT,ntot,2*ntot+1,nextout,nextin,inlist)
         inlist(nextout) = 0          !Don't let v's leave basis
         if (Tshow.eq.1) call disp(T,irdt,icdt,ntot,2*ntot+1,inlist)
      end do

      do i = 1 , m                    !Replace y's with x1's
         nextout = n+i
         nextin = -basicx(i)
         call update (T,irdT,icdT,ntot,2*ntot+1,nextout,nextin,inlist)
         if (Tshow.eq.1) call disp(T,irdt,icdt,ntot,2*ntot+1,inlist)
      end do

      do row = 1 , ntot               !Set v-rows = 0
         if (inlist(row).eq.0) then
            do col = 1 , 2*ntot+1
               T(row,col) = 0
            end do
         end if
      end do

      do col = n+2 , ntot+1           !Set v-cols = 0
         do row = 1 , ntot
            T(row,col) = 0
         end do
      end do

      do col = ntot+2+n , 2*ntot+1    !Set y-cols = 0
         do row = 1 , ntot
            T(row,col) = 0
         end do
      end do

      end if
```

```
C     Is tableau in standard form?

10       do i = 1 , ntot
         do j = i+1 , ntot
            if ((inlist(i).eq.-inlist(j)).and.(inlist(i).ne.0)) then
               std = 'no'
               if (inlist(i).lt.0) then
                  bpairx = i              !bpairx,u=basic pair indices
                  bpairu = j
               else
                  bpairx = j
                  bpairu = i
               end if
               goto 20
            end if
         end do
      end do
      std = 'yes'
20    continue

      iteration = iteration + 1
C     write(1,'(i4,"_")') iteration
      if (iteration.gt.(2*(m+n))) then        !If too many iterations,
         ier=1                                !quit and report error.
         return
      end if

      if (Tshow.eq.1) call disp(T,irdt,icdt,ntot,2*ntot+1,inlist)

C     Non-standard case --

      if (std.eq.'no') then

c     determine nonbasic pair by scanning inlist for missing subscript

      do sub = 1 , ntot
         do j = 1 , ntot
            if (abs(inlist(j)).eq.sub) goto 30
         end do
         nbsub = sub                  !nbsub=missing subscript
         goto 40
30    end do

40    nextin = nbsub                  !This is a u
      us = bpairu                     !Consider when choosing
                                      !outvariable.
```

```
C     Standard case --

      else

c     find smallest basic u

      min=1e10
      do row = 1 , ntot
         if ((T(row,1).lt.min).and.(inlist(row).gt.0)) then
            min = T(row,1)
            minrow = row
         end if
      end do
      if (min.ge.0) goto 999              !Done.

      nextin = -inlist(minrow)            !This is an x
      us = minrow                         !Consider when choosing outvariable

      end if

C     Choose out-variable (variable to leave basis) --

      min=1e10
      if(nextin.gt.0) then
         col = nextin + 1
      else
         col = ntot + 1 + abs(nextin)
      end if
      do row = 1 , ntot
         if((inlist(row).eq.us).or.(inlist(row).lt.0)) then
            di = T(row,1)
            wi = T(row,col)
            diwi = di/wi
            if ((diwi.lt.min).and.(diwi.ge.0).and.(wi.ne.0)) then
               min = diwi
               minrow = row
            end if
         end if
      end do

      nextout = minrow
      call update (T,irdT,icdT,ntot,2*ntot+1,nextout,nextin,inlist)
      goto 10

C     Done --

999   return
      end
```

```
      subroutine disp(T,irdT,icdT,nrows,ncols,inlist)
      implicit none
      integer irdT,icdT,row,col,nrows,ncols,inlist(22)
      real T(irdT,icdT)

c     displays tableau

      do row = 1 , nrows
         write(1,'(i3,"_")') inlist(row)
         do col = 1 , ncols-1
            write(1,'(f5.0,"_")') T(row,col)
         end do
         write(1,'(f5.0)') T(row,col)
      end do
      write(1,*) (inlist(row),row=1,nrows)
      write(1,*) (t(row,1),row=1,nrows)
      write(1,*)

      return
      end
```

```
C-----------------------------------------------------------------------------------------------
      SUBROUTINE UPDATE(T, IRDT, ICDT, NROWS, NCOLS, PIVROW,
                                                          PIVCOL, INLIST)
C-----------------------------------------------------------------------------------------------
C
C     Updates a tableau

      implicit none
      integer irdT,icdT,nrows,ncols,pivrow,pivcol,inlist(irdT),row,col
      real   T(irdT,icdT),pivot


      inlist(pivrow) = pivcol


c     PIVCOL is actually code for variable name for pivot column
c     It must be converted to column #!

      if (pivcol.lt.0) then
         pivcol = nrows + 1 + abs(pivcol)
      else
         pivcol = pivcol + 1
      end if

      pivot = T(pivrow,pivcol)
      if (pivot.ne.0) then
         pivot = 1.0/T(pivrow,pivcol)
      else
         pivot = 1.0
      end if

      do row = 1 , nrows
         if (row.ne.pivrow) then
            T(row,pivcol) = T(row,pivcol)*pivot
            do col = 1 , ncols
               if (col.ne.pivcol) then
               T(row,col)=T(row,col)-T(pivrow,col)*T(row,pivcol)
               end if
            end do
         end if
      end do
      do col = 1 , ncols
         if (col.ne.pivcol) T(pivrow,col) = T(pivrow,col)*pivot
      end do
      do row = 1 , nrows
         T(row,pivcol) = 0
      end do
```

```
if (pivot.ne.0) then
   T(pivrow,pivcol) = 1
else
   T(pivrow,pivcol) = 0
end if

return
end
```

Appendix C:
Multinomial distribution

The multinomial distribution, for an N-dimensional random vector $\mathbf{n}$, characterized by N_t trials is given by

$$p(\mathbf{n}) = N_t! \prod_{i=1}^{N} \frac{p_i^{n_i}}{n_i!} \quad , \tag{C.1}$$

where p_i is the probability of obtaining the ith outcome in one trial.

The covariance of the variables in $\mathbf{n}$ can be determined in the following way. Define a variable $z_{i,p}$ such that

$$z_{i,p} = \begin{cases} 1 & ;\text{outcome } i \text{ is the result of } p\text{th trial} \\ 0 & ;\text{otherwise} \end{cases} \quad , \tag{C.2}$$

and rewrite the variables in $\mathbf{n}$ as

$$n_i = \sum_{p=1}^{N_t} z_{i,p} \quad . \tag{C.3}$$

The covariance of the n-variables is then given by

$$\begin{aligned} \operatorname{cov}(n_i, n_j) = \sigma_{ij} &= \operatorname{cov}\left(\sum_{p=1}^{N_t} z_{i,p} \, , \sum_{q=1}^{N_t} z_{j,q} \right) \\ &= \sum_{p=1}^{N_t} \sum_{q=1}^{N_t} \operatorname{cov}\left(z_{i,p} \, , \, z_{j,q} \right) \quad . \end{aligned} \tag{C.4}$$

Terms having $p \neq q$ are equal to zero since the trials are independent, therefore

$$\sigma_{ij} = \sum_{p=1}^{N_t} \operatorname{cov}\left(z_{i,p} \, , \, z_{j,p} \right) \tag{C.5}$$

$$= \sum_{p=1}^{N_t} \left(< z_{i,p} z_{j,p} > - < z_{i,p} >< z_{j,p} > \right) \quad . \tag{C.6}$$

The correlation term in Eq. (C.6) is equal to zero when $i \neq j$ since there can only be one outcome per trial, therefore σ_{ij} becomes

$$\begin{aligned}\sigma_{ij} &= \sum_{p=1}^{N_t}\left(p_i\delta_{ij} - p_i p_j\right) \\ &= N_t p_i\left(\delta_{ij} - p_j\right) \quad .\end{aligned} \tag{C.7}$$

Quantum-Limited and Classifical-Intensity Image Classification

Miles N.Wernick